4 Multiple Imputation Methods Under MAR

As discussed in Chapter 2, the idea of "filling in" missing values and then applying methods that would have been used with the full data to carry out an analysis under a specified model for the full data has great practical appeal. However, as noted in that chapter, simple, ad hoc strategies in this spirit can lead to compromised inference. A principled approach for which such ***imputation*** of missing data can be justified more rigorously and that yields estimators of precision that take into account the uncertainty associated with imputation is required.

In this chapter, we consider methods for inference in the presence of missing data via ***multiple imputation*** when it is reasonable to assume that the missing data mechanism is MAR. These methods are an alternative to the likelihood-based methods under MAR in Chapter 3.

Throughout this chapter, then, we assume that MAR holds without comment.

4.1 Preliminaries

As in Chapter 3, we assume that interest focuses on a parameter θ (or on components of θ) in a postulated model for the full data $p_Z(z;\theta)$. As reviewed in Section 3.1, if we had a sample of full data, $\underset{\sim}{Z}$, from N individuals, we could obtain the MLE $\widehat{\theta}^F$ for θ by maximizing the corresponding likelihood for the full data or, equivalently, solving in θ the score equation

$$\sum_{i=1}^{N} S_\theta^F(Z_i;\theta) = 0. \tag{4.1}$$

By standard large sample theory, as presented in (3.7), we have that

$$\widehat{\theta}^F \mathrel{\dot\sim} \mathcal{N}[\theta_0, N^{-1}\{N^{-1} I^F(\underset{\sim}{Z};\widehat{\theta}^F)\}^{-1}] = \mathcal{N}[\theta_0, \{I^F(\underset{\sim}{Z};\widehat{\theta}^F)\}^{-1}], \tag{4.2}$$

where

$$I^F(\underset{\sim}{Z};\theta) = -\sum_{i=1}^{N} \frac{\partial^2}{\partial\theta\,\partial\theta^T} \log\{p_Z(Z_i;\theta)\} = -\sum_{i=1}^{N} \frac{\partial}{\partial\theta^T} S_\theta^F(Z_i;\theta). \tag{4.3}$$

Thus, by (4.2), the ***full data observed information matrix*** in (4.3) with $\widehat{\theta}^F$ substituted is an approximation to the covariance matrix of the sampling distribution of $\widehat{\theta}^F$ and characterizes the uncertainty in the full data MLE.

Consider again the ***selection model*** factorization of the joint density of (R, Z) in (3.8), which under the chosen full data model we can write as in (3.9), namely,

$$p_{R,Z}(r, z; \theta, \psi) = p_{R|Z}(r|z; \psi)p_Z(z; \theta).$$

As discussed in Chapter 3, under the ***separability condition*** and MAR, we have ***ignorability*** of the missingness mechanism. Under these conditions, we showed in (3.36) that

$$p_{Z|R,Z_{(R)}}(z|r, z_{(r)}; \theta, \psi) = p_{Z|R,Z_{(R)}}(z|r, z_{(r)}; \theta) = p_{Z|Z_{(r)}}(z|z_{(r)}; \theta) = p_{Z_{(\bar{r})}|Zr}(z_{(\bar{r})}|z_{(r)}; \theta); \tag{4.4}$$

in particular, what we referred to in that chapter as the (frequentist) ***predictive distribution*** depends only on θ and not on ψ. This will be important in the next section.

The idea behind multiple imputation is that one imputes missing data using (4.4) and then estimates θ based on the imputed data and full data methods.

4.2 Fundamentals of multiple imputation

The idea of multiple imputation is due to Don Rubin, who proposed it in several papers in the late 1970s and then fleshed out the methodology in a book in the context of nonresponse in surveys (Rubin, 1987). Since then, there has been an extensive body of work devoted to multiple imputation, including a review by Rubin (1996). Here, we present informally the basic premise. In subsequent sections, we discuss approaches to implementation in more detail.

BASIC IDEA OF MULTIPLE IMPUTATION: Suppose we have a sample of observed data $(R_i, Z_{(R_i)i})$, $i = 1, \ldots, N$, and a full data model $p_Z(z; \theta)$, and the goal is inference on θ. Multiple imputation involves three basic steps or "tasks" (Rubin's terminology):

1. For each individual i, the missing values are ***filled in***, i.e., ***imputed***, M times to create M "full" data sets.

2. The ***full data analysis*** of interest is carried out on ***each*** of these M imputed data sets.

3. The results of the M analyses are ***combined*** into a single analysis that takes into account the imputation.

To accomplish ***Step 1***, for each individual $i = 1, \ldots, N$, we sample at random from the conditional (predictive) distribution in (4.4),

$$p_{Z|R,Z_{(r)}}(z|R_i, Z_{(R_i)i}; \widehat{\theta}^{(m)}), \tag{4.5}$$

M times, $m = 1, \ldots, M$, where, for each m, $\widehat{\theta}^{(m)}$ is some estimator for θ to be discussed later, to obtain

$$Z_i^{(m)}, \quad i = 1, \ldots, N, \;\; m = 1, \ldots, M.$$

We sometimes write $Z_i^{(m)} = Z_i^{(m)}(\widehat{\theta}^{(m)})$ to emphasize dependence on $\widehat{\theta}^{(m)}$. Note that, because of (4.4), there is no need to model the missingness mechanism,

This sampling based on (4.5) can be interpreted as follows. If $R_i = \underset{\sim}{1} = (1, \ldots, 1)^T$, so that full data are observed for individual i and thus $Z_{(R_i)i} = Z_i$, then

$$Z_i^{(m)} = Z_{(R_i)i} = Z_i;$$

i.e., for each m, the imputed full data for i are the observed full data.

For individuals with some components of the full data missing, sampling from (4.5) results in $Z_i^{(m)}$ containing sampled values in the positions where components of the full data are missing and the observed values in all other positions.

If the full data analysis is to estimate θ using maximum likelihood, to carry out ***Step 2***, we obtain $\widehat{\theta}^{*(m)}$ by solving in θ as in (4.1)

$$\sum_{i=1}^{N} S_\theta^F(Z_i^{(m)}; \theta) = 0, \quad m = 1, \ldots, M. \tag{4.6}$$

In ***Step 3***, the ***multiple imputation estimator*** $\widehat{\theta}^*$ for θ is then given by

$$\widehat{\theta}^* = M^{-1} \sum_{m=1}^{M} \widehat{\theta}^{*(m)}, \tag{4.7}$$

the ***average*** of the estimators obtained from each of the M imputed data sets.

RATIONALE: The underlying logic behind multiple imputation is as follows.

If we knew the true value θ_0 of θ, then we could generate random variables $Z_i(\theta_0)$, say, whose distribution has density $p_Z(z;\theta_0)$ by first generating a random $(R_i, Z_{(R_i)i})$ from a distribution with density

$$p_{R,Z_{(R)}}(r, z_{(r)};\theta_0)$$

and then generating a random $Z_i(\theta_0)$ from the conditional distribution

$$p_{Z|R,Z_{(R)}}(z|R_i, Z_{(r)i};\theta_0).$$

For any $i = 1, \ldots, N$, then, $Z_i(\theta_0)$, would be iid with density

$$p_Z(z;\theta_0).$$

The corresponding estimator found by solving in θ

$$\sum_{i=1}^{N} S_\theta^F\{Z_i(\theta_0), \theta) = 0$$

as in (4.1) would be an efficient MLE for the parameter θ.

This is essentially the idea behind multiple imputation; multiple imputation mimics the above data generation process. Assuming that the model is correct, the observed data $(R_i, Z_{(R_i)i})$, $i = 1, \ldots, N$, are generated ***by nature*** from the density $p_{R,Z_{(R)}}(r, z_{(r)};\theta_0)$. However, because we do not know the true value θ_0 of θ, we derive some $\widehat{\theta}^{(m)}$ (to be discussed later) that we believe is a reasonable estimator for θ and then generate $Z_i^{(m)}(\widehat{\theta}^{(m)})$ from the conditional density

$$p_{Z|R,Z_{(R)}}(z|R_i, Z_{(r)i};\widehat{\theta}^{(m)}).$$

4.3 Rubin's variance estimator

The reason that ***multiple*** imputation was proposed by Rubin (instead of creating only one imputed data set) was because the multiple imputations yield an intuitive estimator for the ***sampling variation*** in the imputation estimator $\widehat{\theta}^*$ for θ. Recall from Chapter 2 that ad hoc imputation approaches are generally applied as if the imputed data are observed full data and the usual formulæ for standard errors are used, so that there is no accounting for the imputation in assessment of uncertainty.

To make inference on θ, we require an approximation to the sampling distribution of $\widehat{\theta}^*$. If

$$N^{1/2}(\widehat{\theta}^* - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma^*), \tag{4.8}$$

then the estimator for Σ^* proposed by Rubin is given by

$$\widehat{\Sigma}^* = M^{-1} \sum_{m=1}^{M} \left\{ -N^{-1} \sum_{i=1}^{N} \frac{\partial}{\partial \theta^T} S_\theta^F(Z_i^{(m)}; \widehat{\theta}^{*(m)}) \right\}^{-1} + \left(\frac{M+1}{M} \right) (M-1)^{-1} \sum_{m=1}^{M} N(\widehat{\theta}^{*(m)} - \widehat{\theta}^*)(\widehat{\theta}^{*(m)} - \widehat{\theta}^*)^T. \tag{4.9}$$

Now (4.8) implies that

$$\widehat{\theta}^* \dot{\sim} \mathcal{N}(\theta_0, N^{-1}\widehat{\Sigma}^*), \tag{4.10}$$

so that, from (4.9), an estimator for the sampling covariance matrix of $\widehat{\theta}^*$ is

$$N^{-1}\widehat{\Sigma}^* = M^{-1} \sum_{m=1}^{M} \left\{ -\sum_{i=1}^{N} \frac{\partial}{\partial \theta^T} S_\theta^F(Z_i^{(m)}; \widehat{\theta}^{*(m)}) \right\}^{-1} + \left(\frac{M+1}{M} \right) (M-1)^{-1} \sum_{m=1}^{M} (\widehat{\theta}^{*(m)} - \widehat{\theta}^*)(\widehat{\theta}^{*(m)} - \widehat{\theta}^*)^T. \tag{4.11}$$

The expression in (4.11) has intuitive appeal.

- The first term is the average of estimators for the full data approximate covariance matrices of the $\widehat{\theta}^{*(m)}$ using the inverses of the full data observed information matrices, as in (4.3).
- The second term is the sample covariance matrix of the $\widehat{\theta}^{*(m)}$, $m = 1, \ldots, M$, multiplied by a finite M ***correction factor***.

The expression (4.11) (or equivalently (4.9)) is referred to as ***Rubin's variance estimator***. Standard errors for the components of $\widehat{\theta}^*$ can be obtained from (4.11) in the usual way.

Thus, the ***appeal*** of multiple imputation is that, if the full data estimator for θ and a large sample approximation to its sampling covariance matrix are easy to compute, or if off-the-shelf software for doing so exists, then inference on θ when some data are missing is also "easy." Moreover, the estimator and the form of its approximate covariance estimator are intuitive.

4.4 Proper versus improper imputation

There are two types of imputation schemes that can be used to implement ***Step 1***, ***improper*** and ***proper*** imputation.

IMPROPER IMPUTATION: With ***improper imputation***, we start with an ***initial estimator*** for θ, $\widehat{\theta}^{(init)}$, say, that is obtained in some fashion from the observed data and that is ***consistent and asymptotically normal***, i.e.,

$$N^{1/2}(\widehat{\theta}^{(init)} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\{0, \Sigma^{(init)}(\theta_0)\}. \tag{4.12}$$

In ***Step 1***, for each $m = 1, \ldots, M$, we draw

$$Z_i^{(m)}(\widehat{\theta}^{(init)}), \quad i = 1, \ldots, N$$

from the conditional distributions with conditional densities

$$p_{Z|R,Z_{(R)}}(z|R_i, Z_{(R_i)i}; \widehat{\theta}^{(init)}), \quad i = 1, \ldots, N.$$

Note that $\widehat{\theta}^{(init)}$ is used for all $m = 1, \ldots, M$.

The multiple imputation estimator $\widehat{\theta}^*$ is then obtained following (4.6) by solving in θ

$$\sum_{i=1}^{N} S_\theta^F\{Z_i^{(m)}(\widehat{\theta}^{(init)}); \theta\} = 0, \quad m = 1, \ldots, M$$

to yield $\widehat{\theta}^{*(m)}$ (***Step 2***), $m = 1, \ldots, M$, and substituting the $\widehat{\theta}^{*(m)}$ in (4.7). (***Step 3***).

PROPER OR BAYESIAN IMPUTATION: With ***proper*** or ***Bayesian imputation***, the $Z_i^{(m)}$ are generated taking a Bayesian perspective. From a Bayesian point of view, the parameter θ is regarded as random. Accordingly, the Bayesian approach to ***Step 1*** is to sample $Z_i^{(m)}$, $m = 1, \ldots, M$, for individual i from the ***posterior*** (***Bayesian***) ***predictive distribution***

$$p_{Z|R,Z_{(R)}}(z|R_i, Z_{(R_i)i}) = \int p_{Z|R,Z_{(R)}}(z|R_i, Z_{(R_i)i}; \theta)\, p_{\theta|R,Z_{(R)}}(\theta|R_i, Z_{(R_i)i})\, d\nu(\theta). \tag{4.13}$$

In (4.13), $p_{\theta|R,Z_{(R)}}(\theta|R, Z_{(R)})$ is the ***posterior density*** of θ given the observed data.

From (4.13), then, to implement proper imputation in ***Step 1***, we need to specify a ***prior distribution*** for θ from which the posterior distribution for θ can be obtained using ***Bayes' rule***. Specifically, given a prior density $p_\theta(\theta)$ for θ,

$$p_{\theta|R,Z_{(R)}}(\theta|R, Z_{(R)}) = \frac{p_{R,Z_{(R)}}(R, Z_{(R)}; \theta)\, p_\theta(\theta)}{\int p_{R,Z_{(R)}}(R, Z_{(R)}; \theta)\, p_\theta(\theta)\, d\nu(\theta)}.$$

Step 1 can then be implemented as follows. For each individual $i = 1, \ldots, N$, for each $m = 1, \ldots, M$

(i) Generate $\theta^{(m)}$ from the posterior distribution

$$p_{\theta|R,Z_{(R)}}(\theta|R_i, Z_{(R_i)i}).$$

(ii) Conditional on $\theta^{(m)}$, obtain $Z_i^{(m)}(\theta^{(m)})$ by sampling from

$$p_{Z|R_i,Z_{(R)}}(z|R_i, Z_{(R_i)i}; \theta^{(m)}).$$

The resulting $Z_i^{(m)}(\theta^{(m)})$, $i = 1, \ldots, N$, $m = 1, \ldots, M$, are draws from the posterior predictive distribution in (4.13).

As with improper imputation, the multiple imputation estimator $\widehat{\theta}^*$ is then obtained by solving in θ

$$\sum_{i=1}^{N} S_\theta^F\{Z_i^{(m)}(\theta^{(m)}); \theta\} = 0, \quad m = 1, \ldots, M,$$

to yield $\widehat{\theta}^{*(m)}$ (***Step 2***), $m = 1, \ldots, M$, and substituting the $\widehat{\theta}^{*(m)}$ in (4.7). (***Step 3***).

REMARKS:

- In improper imputation, then, ***Step 1*** is carried out by ***fixing*** θ at some initial estimator $\widehat{\theta}^{(init)}$, whereas in proper imputation, ***Step 1*** involves ***sampling*** θ from its posterior distribution given the observed data.
- Both proper and improper multiple imputation estimators lead to consistent, asymptotically normal estimators $\widehat{\theta}^*$ for θ.
- As we will see in the next section, Rubin's variance estimator $\widehat{\Sigma}^*$ given by (4.9) is a consistent estimator for the asymptotic covariance matrix of a ***proper*** multiple imputation estimator for θ but will be an ***underestimate*** for an ***improper*** multiple imputation estimator.
- The logic behind Bayesian proper imputation seems a bit ***circular***. In general, under the Bayesian paradigm, under suitable regularity conditions and choice of prior distribution for θ, the posterior mean or mode of θ is generally an ***efficient*** estimator for θ.

  Using proper imputation, we draw from the posterior distribution of θ, the mean of which is already an efficient estimator for θ. After imputing M data sets and carrying out ***Steps 2*** and ***3***, we end up with an estimator for θ that is ***no longer*** efficient, although it can be shown that the loss of efficiency goes to zero as M increases. We discuss this further in the next section.

4.5 Asymptotic results

Important insights regarding the differences between improper and proper imputation can be gained by examining the properties of the resulting estimators for θ from a large-sample (frequentist) point of view. In this section, we state several results without proof; much of the development behind these results is due to Wang and Robins (1998) and Robins and Wang (2000) and is presented in detail in Chapter 14 of Tsiatis (2006).

As reviewed in Section 3.1, if we had full data, the MLE $\widehat{\theta}^F$ for θ is fully efficient, with asymptotic covariance matrix $\{\mathcal{I}^F(\theta_0)\}^{-1}$, where $\mathcal{I}^F(\theta_0)$ is the full data expected information matrix. In contrast, with observed data, as discussed in Section 3.3, from (3.29), the MLE $\widehat{\theta}$ is the efficient observed data estimator, with asymptotic covariance matrix $\{\mathcal{I}(\theta_0)\}^{-1}$, where $\mathcal{I}(\theta_0)$ is the observed data expected information matrix.

Of course, the full data MLE is relatively more efficient than the observed data MLE, because there is more information available with full data than there is when some data are missing. This is reflected by the fact that

$$\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0) \tag{4.14}$$

is ***nonnegative definite*** or, equivalently, with $\mathcal{I}^F(\theta_0)$ and $\mathcal{I}(\theta_0)$ both positive definite, that

$$\{\mathcal{I}(\theta_0)\}^{-1} - \{\mathcal{I}^F(\theta_0)\}^{-1}$$

is nonnegative definite. From Section 3.4, that the expression (4.14) is nonnegative definite follows from the ***missing information principle*** presented in (3.68), and (4.14) is often called the ***missing information***, as it reflects the information that is lost due to missing data.

We now consider the large sample properties of imputation estimators.

ASYMPTOTIC RESULTS FOR IMPROPER IMPUTATION ESTIMATOR: For improper imputation using an initial estimator $\widehat{\theta}^{(init)}$ with asymptotic covariance matrix $\Sigma^{(init)}(\theta_0)$ as in (4.12), denote the final estimator for θ from ***Step 3*** by $\widehat{\theta}^{*(improper)}$.

Then it can be shown that

$$N^{1/2}(\widehat{\theta}^{*(improper)} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma^{*(improper)}),$$

where

$$\begin{aligned}\Sigma^{*(improper)} &= \{\mathcal{I}^F(\theta_0)\}^{-1} + \left(\frac{M+1}{M}\right)\{\mathcal{I}^F(\theta_0)\}^{-1}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}\{\mathcal{I}^F(\theta_0)\}^{-1} \\ &\quad + \{\mathcal{I}^F(\theta_0)\}^{-1}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}\Sigma^{(init)}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}\{\mathcal{I}^F(\theta_0)\}^{-1} \qquad (4.15)\end{aligned}$$

Under these conditions, the large sample behavior of Rubin's variance estimator $\widehat{\Sigma}^*$ given in (4.9) can also be deduced. Here, both terms in (4.9) are based on using $Z_i^{(m)} = Z_i^{(m)}(\widehat{\theta}^{(init)})$. Then Rubin's variance estimator $\widehat{\Sigma}^*$ converges as $N \to \infty$ to

$$\Sigma^{*(Rubin)} = \{\mathcal{I}^F(\theta_0)\}^{-1} + \left(\frac{M+1}{M}\right)\{\mathcal{I}^F(\theta_0)\}^{-1}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}\{\mathcal{I}^F(\theta_0)\}^{-1}, \qquad (4.16)$$

which is equal to the first two terms in the expression for $\Sigma^{*(improper)}$ in (4.15). Thus, comparing (4.16) to (4.15), it is clear that Rubin's variance estimator used with improper imputation ***underestimates*** the true asymptotic variance of the improper imputation estimator $\widehat{\theta}^{*(improper)}$.

Accordingly, if one were to carry out improper imputation and use Rubin's variance estimator, the resulting inferences would be ***optimistic***, failing to account faithfully for the true extent of uncertainty involved.

SORT-OF PROPER IMPUTATION: Another strategy for multiple imputation that can be viewed as somewhere between improper and fully Bayesian proper imputation is as follows.

Instead of fixing $\widehat{\theta}^{(init)}$ for each $m = 1, \ldots, M$, suppose that we randomly draw $\widehat{\theta}^{(m)}$ from a

$$\mathcal{N}(\widehat{\theta}^{(init)}, \widehat{\Sigma}^{(init)}) \qquad (4.17)$$

distribution, where $\widehat{\Sigma}^{(init)}$ is a consistent estimator for the true asymptotic covariance matrix $\Sigma^{(init)}$ of $\widehat{\theta}^{(init)}$. By doing so, we are roughly drawing from the posterior distribution in an asymptotic sense; for large N and certain choices of prior distribution for θ, there is a correspondence between the posterior and asymptotic distributions.

We refer to the resulting multiple imputation estimator from ***Step 3*** as $\widehat{\theta}^{*(proper)}$.

It can be shown that

$$N^{1/2}(\widehat{\theta}^{*(proper)} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma^{*(proper)}),$$

where

$$\begin{aligned}\Sigma^{*(proper)} &= \{\mathcal{I}^F(\theta_0)\}^{-1} + \left(\frac{M+1}{M}\right)\{\mathcal{I}^F(\theta_0)\}^{-1}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}\{\mathcal{I}^F(\theta_0)\}^{-1} \\ &\quad + \left(\frac{M+1}{M}\right)\{\mathcal{I}^F(\theta_0)\}^{-1}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}\Sigma^{(init)}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}\{\mathcal{I}^F(\theta_0)\}^{-1}. \quad (4.18)\end{aligned}$$

Comparing (4.18) to (4.15) shows that the asymptotic covariance matrix of $\widehat{\theta}^{*(proper)}$ is actually ***larger*** in the sense of nonnegative definiteness than that of $\widehat{\theta}^{*(improper)}$, although the difference converges to zero as $M \to \infty$.

Under these conditions, it turns out that Rubin's variance estimator (4.9) converges to $\Sigma^{*(proper)}$, so that it is a valid estimator for the asymptotic covariance matrix of the (***sort-of***) proper imputation estimator $\widehat{\theta}^{*(proper)}$. This is in contrast to its use with improper imputation, where it is an underestimate, as presented above.

REMARKS:

- It is possible to derive consistent estimators for the asymptotic covariance matrix of the improper imputation estimator $\widehat{\theta}^{*(improper)}$; see Chapter 14 of Tsiatis (2006). However, these have forms that are more difficult than that of Rubin's estimator and are thus not appealing for practical use.
- If the initial estimator $\widehat{\theta}^{(init)}$ is itself asymptotically efficient; e.g., if $\widehat{\theta}^{(init)}$ is the observed data MLE for θ or the posterior mean of θ given the observed data, then

  $$N^{1/2}(\widehat{\theta}^{(init)} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}[0, \{\mathcal{I}(\theta_0)\}^{-1}].$$

  It can be shown that, substituting $\{\mathcal{I}(\theta_0)\}^{-1}$ for $\Sigma^{(init)}$ in the expression for $\Sigma^{*(proper)}$ in (4.18),

  $$\Sigma^{*(proper)} = \{\mathcal{I}(\theta_0)\}^{-1} + M^{-1}[\{\mathcal{I}(\theta_0)\}^{-1} - \{\mathcal{I}^F(\theta_0)\}^{-1}] \quad (4.19)$$

  (try it). In this case, (4.19) shows that carrying out proper multiple imputation starting with an efficient initial estimator entails a ***loss of efficiency*** reflected by the term

  $$M^{-1}[\{\mathcal{I}(\theta_0)\}^{-1} - \{\mathcal{I}^F(\theta_0)\}^{-1}]$$

  relative to having just used the initial estimator.

  This seems a bit ***odd*** but is what proper Bayesian imputation implies.

INEFFICIENT INITIAL ESTIMATOR: What happens if instead we start with an inefficient but possibly easy-to-compute estimator $\widehat{\theta}^{(init)}$?

Under these conditions, it can be shown that the resulting improper or proper multiple imputation estimator is ***asymptotically equivalent*** to the ***one-step updated EM algorithm estimator***.

To see what we mean by this, recall from Section 3.3 that the observed data MLE $\widehat{\theta}$ can be expressed as the solution to the observed data score equation given in (3.38); namely, $\widehat{\theta}$ satisfies

$$\sum_{i=1}^{N} S_\theta(R_i, Z_{(R_i)i}; \widehat{\theta}) = \sum_{i=1}^{N} E_{\widehat{\theta}}\Big\{ S_\theta^F(Z_i; \widehat{\theta}) | R_i, Z_{(R_i)i} \Big\} = 0, \tag{4.20}$$

where the equality in (4.20) follows from (3.33).

The EM algorithm is an iterative process for finding the MLE and thus solving (4.20). In Section 3.4, we characterized the EM in the classical way as an iterative maximization, but it can be cast equivalently as an iterative process involving solving the observed data score equation. That is, if $\theta^{(t)}$ is the tth iterate, then the $(t+1)$th iterate satisfies

$$\sum_{i=1}^{N} E_{\theta^{(t)}}\Big\{ S_\theta^F(Z_i; \theta^{(t+1)}) | R_i, Z_{(R_i)i} \Big\} = 0. \tag{4.21}$$

Ordinarily, we would iterate (4.21) to convergence to obtain the observed data MLE, but we can consider each successive iterate itself as an ***estimator*** for θ.

Thus, by ***one-step updated EM algorithm estimator***, we mean the estimator $\widehat{\theta}^{(em,1)}$, say, satisfying

$$\sum_{i=1}^{N} E_{\widehat{\theta}^{(init)}}\Big\{ S_\theta^F(Z_i; \widehat{\theta}^{(em,1)}) | R_i, Z_{(R_i)i} \Big\} = 0.$$

In fact, writing the tth iterate as an estimator $\widehat{\theta}^{(em,t)}$, because as $t \to \infty$, the EM algorithm converges (under regularity conditions) to the observed data MLE, which is asymptotically efficient, we have that

$$N^{1/2}(\widehat{\theta}^{(em,t)} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma^{(em,t)}), \quad \text{where} \quad \Sigma^{(em,t)} \to \{\mathcal{I}(\theta_0)\}^{-1} \ \text{ as } \ t \to \infty. \tag{4.22}$$

In Theorem 14.5 of Tsiatis (2006), it is shown that the asymptotic distribution of the multiple imputation estimator as $M \to \infty$ is ***the same*** as that of the one-step EM algorithm estimator. The result in (4.22) thus suggests another multiple imputation strategy. After M imputations using $\widehat{\theta}^{(init)}$ (improper or proper) to obtain the imputation estimator $\widehat{\theta}^{*(1)}$, say, ***restart*** the imputation process, now using $\widehat{\theta}^{*(1)}$ as the initial estimator in the imputations. By continuing this process, theoretically, as suggested by (4.22), we can iterate toward a fully efficient imputation estimator.

To implement such a strategy, M must be chosen sufficiently large to ensure that each successive iterate is relatively more efficient than the previous one. As the process converges toward efficiency, M must be made increasingly larger.

SUMMARY: After this discussion of the properties of multiple imputation estimators, this general approach may not seem as attractive as it may have at the outset. However, from a practical point of view, multiple imputation has some nice features, discussed in the next several sections.

4.6 Imputation from a multivariate normal distribution

IMPUTER'S MODEL VERSUS ANALYST'S MODEL: Our development up to now has assumed that the model that is used to draw the imputations is ***the same*** as that to be used to analyze the data. The term "***congenial***," coined by Meng (1994), has been used to describe the situation where the models used for imputation and analysis are ***compatible*** in this sense.

However, for general problems, it may not always be feasible or convenient to use congenial models, because it may be difficult to generate imputations in the context of a complex full data model. This has led to examination of performance of multiple imputation methods when the models are ***not congenial***; that is, as it is often portrayed in the literature, when the "***imputer's model***" differs from the "***analyst's model***."

In particular, suppose that imputation is carried out assuming that the full data are ***multivariate normally distributed***. Specifically, for $Z = (Z_1, \ldots, Z_K)$, where each component Z_k, $k = 1, \ldots, K$, is of dimension p_k, say, if we interpret Z as a vector of dimension $(P \times 1)$, $P = \sum_{k=1}^{K} p_k$, then we mean that we assume Z is P-variate normal and treat each scalar component as either observed or missing.

It turns out, not surprisingly, that under this condition imputation of missing data is relatively easily to implement, as we demonstrate shortly. Here, then, we impute missing data $Z_i^{(m)}$ by generating from the conditional density

$$p_{Z|R,Z_{(R)}}(z|R_i, Z_{(R_i)i}, \xi)$$

that arises from assuming (possibly wrongly) that Z is a multivariate normal as described above, indexed by mean and covariance parameters ξ. Of course, if the model for the full data $p_Z(z;\theta)$ assumed by the analyst is not multivariate normal, then there is nothing to suggest that the resulting multiple imputation estimator $\widehat{\theta}^*$ need be a "***good***" estimator for θ.

Interestingly, it has been shown via extensive numerical studies that, if the proportion of missing data is not great, this approach can lead to reasonable results, even if missing components of Z are not continuous but are discrete/categorical. Schafer (1997) offers extensive discussion.

Accordingly, available software for multiple imputation, such as SAS `proc mi` and R packages such as `norm`, `amelia`, or `mice`, base imputation fully or in part on a multivariate normal model.

We demonstrate how imputation can be carried out assuming multivariate normality first in the case where missingness is ***monotone***.

MONOTONE MISSINGNESS: For simplicity, consider the simple longitudinal situation where a scalar outcome Y_j is collected at times t_j, $j = 1, \ldots, T$, and assume that the full data are

$$Z = (Y_1, \ldots, Y_T). \tag{4.23}$$

As usual, let $D = j + 1$ correspond to dropout at time t_{j+1}, so that $(Y_1, \ldots, Y_j)$ is observed and $(Y_{j+1}, \ldots, Y_T)$ is missing, $j = 1, \ldots, T$, and thus $Z_{(D)} = (Y_1, \ldots, Y_j)$; and $D = T + 1$ means $Z_{(D)} = Z$.

Suppose that, for the purpose of imputation, we assume that Z (interpreted as the $(T \times 1)$ vector $Y = (Y_1, \ldots, Y_T)^T$ as above) is distributed as multivariate normal. We emphasize that the full data model $p_Z(z; \theta)$ for Z of interest may ***not*** be multivariate normal.

A T-variate normal distribution can be specified fully in terms of a $(T \times 1)$ mean vector and $T(T + 1)/2$ distinct variance and covariance parameters, for a total of $T + T(T + 1)/2$ parameters. For our purposes, it is convenient to parameterize the T-variate normal instead in terms of a different set of $T + T(T + 1)/2$ parameters as follows. If we take

$$\begin{aligned}
Y_1 &\sim \mathcal{N}(\alpha_{01}, \sigma_1^2) \\
&\vdots \\
Y_j | Y_1, \ldots, Y_{j-1} &\sim \mathcal{N}(\alpha_{0j} + \alpha_{1j} Y_1 + \cdots + \alpha_{j-1,j} Y_{j-1}, \sigma_j^2) \\
&\vdots \\
Y_T | Y_1, \ldots, Y_{T-1} &\sim \mathcal{N}(\alpha_{0T} + \alpha_{1T} Y_1 + \cdots + \alpha_{T-1,T} Y_{T-1}, \sigma_T^2),
\end{aligned} \tag{4.24}$$

then it follows that Y is multivariate normal with mean vector and covariance matrix determined by the $T(T + 1)/2 + T$ parameters in

$$\xi = (\overline{\alpha}_1^T, \ldots, \overline{\alpha}_T^T, \sigma_1^2, \ldots, \sigma_T^2)^T,$$

where $\overline{\alpha}_j = (\alpha_{0j}, \alpha_{1j} \ldots, \alpha_{j-1,j})^T$, $j = 1, \ldots, T$ (convince yourself).

To facilitate imputation, the multivariate normal model in (4.24) is first fitted to the observed data $(D_i, Z_{(D_i)i})$, $i = 1, \ldots, N$, by maximum likelihood under MAR to estimate ξ, and then the estimates are used in a scheme to generate imputed data; here, we describe a proper imputation approach.

To obtain the MLE for ξ under the (***imputer's***) model (4.24), note that the likelihood contribution for a single observation $(D, Z_{(D)})$ under MAR, suppressing the i subscript, can be written as

$$\prod_{q=1}^{T} \left\{ p_{Y_1,\ldots,Y_q}(Y_1, \ldots, Y_q; \overline{\alpha}_1, \ldots, \overline{\alpha}_q, \sigma_1^2, \ldots, \sigma_q^2) \right\}^{I(D=q+1)} \tag{4.25}$$

$$= \prod_{q=1}^{T} \left\{ \prod_{j=1}^{q} p_{Y_j|Y_1,\ldots,Y_{j-1}}(Y_j|Y_1, \ldots, Y_{j-1}; \overline{\alpha}_j, \sigma_j^2) \right\}^{I(D=q+1)}, \tag{4.26}$$

where the densities in (4.25) are multivariate normal as indicated, and the conditional densities in (4.26) correspond to the conditional normal specifications in (4.24) .

Interchanging the order of the products in (4.26), we obtain

$$\prod_{j=1}^{T} \left\{ \prod_{q=j}^{T} p_{Y_j|Y_1,\ldots,Y_{j-1}}(Y_j|Y_1, \ldots, Y_{j-1}; \overline{\alpha}_j, \sigma_j^2) \right\}^{I(D=q+1)}$$

$$= \prod_{j=1}^{T} \left\{ p_{Y_j|Y_1,\ldots,Y_{j-1}}(Y_j|Y_1, \ldots, Y_{j-1}; \overline{\alpha}_j, \sigma_j^2) \right\}^{I(D\geq j+1)}. \tag{4.27}$$

Thus, from (4.27), for a sample of iid data $(D_i, Z_{(D_i)i})$, $i = 1 \ldots, N$, the likelihood can be written as

$$\prod_{j=1}^{T} \left\{ \prod_{i:D_i\geq j+1} p_{Y_j|Y_1,\ldots,Y_{j-1}}(Y_{ij}|Y_{i1}, \ldots, Y_{i,j-1}; \overline{\alpha}_j, \sigma_j^2) \right\}. \tag{4.28}$$

From (4.28), because $(\overline{\alpha}_j, \sigma_j^2)$ separate in the likelihood for $j = 1, \ldots, T$, the MLEs $(\widehat{\overline{\alpha}}_j, \widehat{\sigma}_j^2)$, $j = 1, \ldots, T$, under this multivariate normal model can be obtained separately for each j by maximizing in $(\overline{\alpha}_j, \sigma_j^2)$

$$\prod_{i:D_i\geq j+1} p_{Y_j|Y_1,\ldots,Y_{j-1}}(Y_{ij}|Y_{i1}, \ldots, Y_{i,j-1}; \overline{\alpha}_j, \sigma_j^2),$$

and, moreover, the estimators for each $j = 1, \ldots, T$ are ***asymptotically independent***.

Because of the formulation (4.24), for each $j = 1, \ldots, T$, $\widehat{\overline{\alpha}}_j$ is the ordinary least squares (OLS) estimator derived using all the individuals i for whom $Y_1, \ldots, Y_j$ are observed.

That is, $\widehat{\overline{\alpha}}_j$ is obtained by fitting by OLS the model

$$Y_{ij} = \alpha_{0j} + \alpha_{1j} Y_{i1} + \cdots + \alpha_{j-1,j} Y_{i,j-1} + \sigma_j \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0,1)$$

to the data for all i for whom $D_i \geq j+1$, where the ϵ_{ij} are iid. The estimator $\widehat{\sigma}_j^2$ is obtained as

$$\widehat{\sigma}_j^2 = (N_j - j)^{-1} \sum_{i:D_i \geq j+1} (Y_{ij} - \widehat{\alpha}_{0j} - \widehat{\alpha}_{1j} Y_{i1} - \cdots - \widehat{\alpha}_{j-1,j} Y_{i,j-1})^2, \quad N_j = \sum_{i=1}^{N} I(D_i \geq j+1).$$

Armed with these results, ***proper imputation*** can be implemented as follows. For each $m = 1, \ldots, M$:

(i) Draw $\xi^{(m)}$ from the ***posterior distribution***. Namely, for $j = 1, \ldots, T$, obtain

$$\sigma_j^{2(m)} = \widehat{\sigma}_j^2 (N_j - j)/q,$$

where q is a random draw from a $\chi^2_{N_j - j}$ distribution, and

$$\overline{\alpha}_j^{(m)} = \widehat{\overline{\alpha}}_j + \sigma_j^{(m)} (S_j^T S_j)^{-1/2} U_j, \quad j = 1, \ldots, T,$$

where S_j is the usual design matrix corresponding to the jth regression implied by (4.24), $U_j \sim \mathcal{N}(0, I_j)$ (iid), and I_j is a $(j \times j)$ identity matrix.

(ii) For each $i = 1, \ldots, N$, impute

$$Z_i^{(m)}(\xi^{(m)}) = \{ Y_{i1}(\xi^{(m)}), \ldots, Y_{iT}(\xi^{(m)}) \},$$

where, for i such that $D_i = j+1$,

- For $q = 1, \ldots, j$, $Y_{iq}(\xi^{(m)}) = Y_{iq}$, the observed outcome
- Otherwise, generate random deviates

$$Y_{i,j+1}^{(m)}(\xi^{(m)}) = \alpha_{0,j+1}^{(m)} + \alpha_{1,j+1}^{(m)} Y_{i1}(\xi^{(m)}) + \cdots + \alpha_{j,j+1}^{(m)} Y_{ij}(\xi^{(m)}) + \sigma_{j+1}^{(m)} \epsilon_{i,j+1}^{(m)},$$

$$Y_{ir}^{(m)}(\xi^{(m)}) = \alpha_{0r}^{(m)} + \alpha_{1r}^{(m)} Y_{i1}(\xi^{(m)}) + \cdots + \alpha_{r-1,r}^{(m)} Y_{i,r-1}^{(m)}(\xi^{(m)}) + \sigma_r^{(m)} \epsilon_{ir}^{(m)}, \quad r = j+2, \ldots, T,$$

where $\epsilon_{ir}^{(m)}$, $r = j+1, \ldots, T$, are iid standard normal; and $Y_{iq}(\xi^{(m)}) = Y_{iq}$, $q = 1, \ldots, j$.

The foregoing scheme implements ***Step 1*** under the assumption of multivariate normality. ***Steps 2*** and ***3*** would then be carried out by implementing the ***intended*** full analysis under ***actual assumed full data model*** on each of the M imputed data sets to obtain $\widehat{\theta}^{*(m)}$, $m = 1, \ldots, M$, and combining the results.

NONMONOTONE MISSINGESS: We now consider how imputation can be carried out assuming multivariate normality for imputation in the more difficult case where the missingness pattern is ***non-monotone***. Here, one can use ***Markov chain Monte Carlo*** (MCMC) methods to impute missing data from the (Bayesian) posterior predictive distribution and thus implement Rubin's fully proper (Bayesian) imputation. Thus, we need a means of making random draws from the posterior predictive distribution given in (4.13).

GIBBS SAMPLING: Gibbs sampling is the most straightforward and well known form of MCMC and can be used for this purpose. We first give a quick, generic overview of the basic procedure and premise, and then describe how Gibbs sampling can be used for proper imputation.

Suppose that a random vector W is partitioned as $(W_1, \ldots, W_K)$. We would like to sample from the distribution of W with density $p_W(w)$, say. The Gibbs sampling technique involves sampling iteratively from the so-called ***full conditional distributions***; i.e., the conditional distributions of each component of W given all the others.

In particular, given the value of W sampled at iteration t, say,

$$W^{(t)} = (W_1^{(t)}, \ldots, W_K^{(t)}),$$

the value at step $t+1$, $W^{(t+1)} = (W_1^{(t+1)}, \ldots, W_K^{(t+1)})$, is obtained by successively drawing from the distributions, in obvious notation

$$\begin{aligned}
W_1^{(t+1)} &\sim p_{W_1|W_2,\ldots,W_K}(w_1|W_2^{(t)}, \ldots, W_K^{(t)}) \\
W_2^{(t+1)} &\sim p_{W_2|W_1,W_3,\ldots,W_K}(w_2|W_1^{(t+1)}, W_3^{(t)}, \ldots, W_K^{(t)}) \\
&\vdots \\
W_K^{(t+1)} &\sim p_{W_K|W_1,\ldots,W_{K-1}}(w_K|W_1^{(t+1)}, \ldots, W_{K-1}^{(t+1)})
\end{aligned}$$

It can be shown that the sequence $\{W^{(t)}, t = 0, 1, 2, \ldots\}$ forms a ***Markov chain*** that, under mild conditions, has ***stationary distribution*** equal to $p_W(w)$; that is,

$$W^{(t)} \xrightarrow{\mathcal{L}} W \quad \text{as} \quad t \to \infty.$$

PROPER IMPUTATION VIA GIBBS SAMPLING: The foregoing formulation can be used to implement proper imputation as follows. We present this first in the ideal case where the imputer's and analyst's models are the same, and then demonstrate in the case when these models need not be the same.

Here, we wish to make random draws for each individual $i = 1, \ldots, N$ from the posterior predictive distribution. That is, if individual i has $R_i = r$, we would like to draw from (4.13). As we showed in Chapter 3, under MAR, from (3.48), we have

$$p_{Z|Z_{(r)}}\{(Z_{(r)i}, z_{(\bar{r})})|Z_{(r)i};\theta\} = p_{Z|R,Z_{(R)}}\{(Z_{(r)i}, z_{(\bar{r})})|R_i = r, Z_{(r)i};\theta\} = p_{Z_{(\bar{r})}|Z_{(r)}}(z_{(\bar{r})}|Z_{(r)i};\theta).$$

Thus, the posterior predictive distribution from which we wish to sample is

$$p_{Z_{(\bar{r})}|Z_{(r)}}(z_{(\bar{r})}|Z_{(r)i}) = p_{Z|R,Z_{(R)}}\{(Z_{(r)i}, z_{(\bar{r})})|R_i = r, Z_{(r)i}\}. \tag{4.29}$$

Drawing directly from (4.29) can be difficult. However, viewing θ as a random vector under the Bayesian point of view, it may be possible to generate random draws successively from

$$p_{Z_{(\bar{r})}|Z_{(r)}}(z_{(\bar{r})}|Z_{(r)i};\theta) \quad \text{and then} \tag{4.30}$$

$$p_{\theta|Z_{(r)},Z_{(\bar{r})}}(\theta|Z_{(r)i}, z_{(\bar{r})}). \tag{4.31}$$

Identifying for individual i

$$W = (Z_{(\bar{r})i}, \theta),$$

if this were possible, we could implement a Gibbs sampling scheme to do this by generating successive iterates; namely, given the tth iterate $(Z^{(t)}_{(r)i}, \theta^{(t)})$, obtain the $(t+1)$th iterate as, from (4.30) and (4.31),

$$\begin{aligned} Z^{(t+1)}_{(\bar{r})i} &\sim p_{Z_{(\bar{r})}|Z_{(r)};\theta}(z_{(r)}|Z_{(r)i};\theta^{(t)}) \\ \theta^{(t+1)} &\sim p_{\theta|Z_{(r)},Z_{(\bar{r})}}(\theta|Z_{(r)i}, Z^{(t+1)}_{(\bar{r})i}). \end{aligned}$$

As in the generic demonstration, the sequence $(Z^{(t)}_{(\bar{r})i}, \theta^{(t)})$ for $t = 0, 1, 2, \ldots$ generated this way would converge in distribution to a random $(Z_{(\bar{r})i}, \theta)$ from the conditional distribution

$$p_{Z_{(\bar{r})i},\theta|Z_{(r)}}(z_{(\bar{r})}, \theta|Z_{(r)i}).$$

Thus, with t sufficiently large, $Z^{(t)}_{(\bar{r})i}$ would be, roughly speaking, a draw from the posterior predictive distribution

$$p_{Z_{(\bar{r})}|Z_{(r)}}(z_{(\bar{r})}|Z_{(r)i})$$

in (4.29).

Accordingly, this technique could be used to implement proper imputation, where, for each $i = 1, \ldots, N$, the draws $Z_{(\bar{r})i}$ obtained after a large number of iterations t used to "fill in" the missing values. To do so, we would need a starting value $\theta^{(0)}$ and a way to generate random draws from (4.30) and (4.31).

PROPER IMPUTATION USING A MULTIVARIATE NORMAL IMPUTER'S MODEL We now demonstrate how this would be implemented when, as above, we assume that the full data are ***multivariate normally distributed***. Again consider the situation where the full data Z are analogous to (4.23),

$$Z = (Y_1, \ldots, Y_K),$$

say, where each Y_k is a real-valued, normally-distributed outcome. This might correspond to a longitudinal situation where $K = T$ or more generally to any situation where K variables ideally are to be observed on all individuals. Suppose that components of Z can be missing in an ***intermittent***, nonmonotone fashion.

Writing $Y = (Y_1, \ldots, Y_K)^T$, then, ***for the purposes of imputation***, we assume

$$Y \sim \mathcal{N}(\mu, \Sigma), \quad \mu = (\mu_1, \ldots, \mu_K)^T, \tag{4.32}$$

and Σ is a $(K \times K)$ covariance matrix. Let ξ denote collectively (μ, Σ).

Consider a fixed $r = (r_1, \ldots, r_K)^T$, and rearrange and partition Y as $(Y_{(r)}^T, Y_{(\bar{r})}^T)^T$, so into observed and unobserved components. For individual i for whom $R_i = r$, under this normal ***imputer's model***, from (4.30) and (4.31), we wish to draw successively from

$$p_{Y_{(\bar{r})}|Y_{(r)};\xi}(y_{(\bar{r})}|Y_{(r)i};\xi) \quad \text{and then} \tag{4.33}$$

$$p_{\xi|Y_{(r)},Y_{(\bar{r})}}(\xi|Y_{(r)i}, y_{(\bar{r})}). \tag{4.34}$$

It turns out that straightforward methods are available for making draws from (4.33) based on the properties of the conditional distributions a multivariate normal distribution and from (4.34) under ***conjugate prior*** distributions, as follows.

Partition μ and Σ analogous to the partition of Y as

$$\mu = (\mu_{(r)}^T, \mu_{(\bar{r})}^T)^T \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{(rr)} & \Sigma_{(r\bar{r})} \\ \Sigma_{(\bar{r}r)} & \Sigma_{(\bar{r}\bar{r})} \end{pmatrix}.$$

Then a well-known result for the multivariate normal distribution is that the ***conditional distribution*** of $Y_{(\bar{r})}$ given $Y_{(r)}$ is also multivariate normal with mean

$$E(Y_{(\bar{r})}|Y_{(r)}) = \mu_{(\bar{r})} + \Sigma_{(\bar{r}r)}\Sigma_{(rr)}^{-1}(Y_{(r)} - \mu_{(r)}), \tag{4.35}$$

and covariance matrix

$$\Sigma_{(\bar{r}\bar{r})} - \Sigma_{(\bar{r}r)}\Sigma_{(rr)}^{-1}\Sigma_{(r\bar{r})}. \tag{4.36}$$

For individual i for whom $R_i = r$, making a random draw from (4.33) involves drawing from the multivariate normal distribution with mean (4.35) and and covariance matrix (4.36). That is, in the Gibbs sampling scheme, given the tth iterate $(Y_{(\bar{r})i}^{(t)}, \xi^{(t)})$, obtain the $(t+1)$th iterate as

$$Y_{(\bar{r})i}^{(t+1)} \sim p_{Y_{(\bar{r})}|Y_{(r)};\xi}(y_{(\bar{r})}|Y_{(r)i};\xi^{(t)}), \tag{4.37}$$

where $p_{Y_{(\bar{r})}|Y_{(r)}}(y_{(\bar{r})}|Y_{(r)i};\xi^{(t)})$ in (4.37) is the multivariate normal density with, from (4.35) and (4.36), mean

$$\mu_{(\bar{r})}^{(t)} + \Sigma_{(\bar{r}r)}^{(t)}\Sigma_{(rr)}^{(t)-1}(Y_{(r)i} - \mu_{(r)}^{(t)})$$

and covariance matrix

$$\Sigma_{(\bar{r}\bar{r})}^{(t)} - \Sigma_{(\bar{r}r)}^{(t)}\Sigma_{(rr)}^{(t)-1}\Sigma_{(r\bar{r})}^{(t)}.$$

We now discuss how to make random draws from (4.34) to obtain the iterate $\xi^{(t)}$. Note that (4.34) is the posterior density of ξ given the full data, where, here, the full data are taken to be multivariate normal as in (4.32).

In ***Bayesian inference***, the simplest way to proceed is to choose a class of ***prior distributions*** that is ***conjugate*** to the likelihood function. A conjugate class has the property that any prior distribution $p_\xi(\xi)$, say, in the class leads to a posterior distribution $p_{\xi|Y}(\xi|y)$ in our case that is also in the class.

NORMAL INVERTED WISHART DISTRIBUTION: When Y is multivariate normal as in (4.32) and both μ and Σ are unknown, the most natural conjugate class is that of the ***normal inverted Wishart distribution***.

The ***Wishart distribution*** is the ***multivariate generalization*** of the chi-square distribution. We now define the Wishart and inverted Wishart distributions generically and then apply the results to the problem at hand. If X is a $(m \times p)$ matrix whose rows are iid $\mathcal{N}(0, \Lambda^{-1})$ for $(p \times p)$ covariance matrix Λ^{-1}, then the matrix of sums of squares

$$A = X^T X$$

is said to have a Wishart distribution, and we write

$$A \sim \mathcal{W}(m, \Lambda^{-1}).$$

The parameter m is referred to as the ***degrees of freedom*** and Λ^{-1} as the ***scale***.

If $A \sim \mathcal{W}(m, \Lambda^{-1})$, then $B = A^{-1}$ is said to have an ***inverted Wishart*** or ***inverse Wishart distribution***, which we write as

$$B \sim \mathcal{W}^{-1}(m, \Lambda). \tag{4.38}$$

With Y multivariate normal as in (4.32) and ξ defined as above, we can define the ***normal inverted Wishart prior and posterior*** for ξ as follows. Recall that we are viewing ξ, and thus μ and Σ, as random. Suppose that, given Σ, μ is assumed to have a multivariate normal prior distribution

$$\mu|\Sigma \sim \mathcal{N}(\mu_0, \tau^{-1}\Sigma), \quad \tau > 0, \tag{4.39}$$

where μ_0 is fixed and known, and

$$\Sigma \sim \mathcal{W}^{-1}(m, \Lambda) \tag{4.40}$$

as in (4.38) for fixed and known m and Λ. Then the resulting joint prior density for ξ can be written as

$$p_\xi(\xi) = p_{\mu,\Sigma}(\mu, \Sigma) = p_{\mu|\Sigma}(\mu|\Sigma)\, p_\Sigma(\Sigma), \tag{4.41}$$

where the densities in the rightmost expression in (4.41) are those of the distributions in (4.39) and (4.40), respectively. Then ξ satisfying (4.41) is said to have a ***normal inverted Wishart distribution***.

If we have iid data $\underset{\sim}{Y} = \{Y_i,\, i = 1, \ldots, N\}$, it can be shown that

$$p_{\xi|\underset{\sim}{Y}}(\xi|\underset{\sim}{Y}) = p_{\mu,\Sigma|\underset{\sim}{Y}}(\mu, \Sigma|\underset{\sim}{Y}) = p_{\mu|\Sigma,\underset{\sim}{Y}}(\mu|\Sigma, \underset{\sim}{Y})\, p_{\Sigma|\underset{\sim}{Y}}(\Sigma|\underset{\sim}{Y}), \tag{4.42}$$

where the densities in (4.42) are such that

$$\mu|\Sigma, \underset{\sim}{Y} \sim \mathcal{N}(\mu_0', (\tau')^{-1}\Sigma), \qquad \Sigma|\underset{\sim}{Y} \sim \mathcal{W}^{-1}(m', \Lambda'),$$

$$\tau' = \tau + N, \quad m' = m + N, \qquad \mu_0' = \left(\frac{N}{\tau + N}\right)\overline{Y} + \left(\frac{\tau}{\tau + N}\right)\mu_0,$$

$$\Lambda' = \left\{\Lambda + NS + \left(\frac{\tau N}{\tau + N}\right)(\overline{Y} - \mu_0)(\overline{Y} - \mu_0)^T\right\},$$

and

$$\overline{Y} = N^{-1}\sum_{i=1}^{N} Y_i, \qquad S = (N-1)^{-1}\sum_{i=1}^{N}(Y_i - \overline{Y})(Y_i - \overline{Y})^T. \tag{4.43}$$

The density in (4.42) is thus that of the posterior distribution given the full sample data that results from adopting the joint prior density in (4.41).

When the result (4.42) is used in practice, one often uses a ***noninformative improper prior*** where one lets $\tau \to 0$, $m \to -1$, and $\Lambda \to 0$. Under these conditions, the posterior distribution of ξ given the full sample data in (4.42) becomes the normal inverted Wishart distribution with the densities in (4.41) corresponding to

$$\mu|\Sigma, \underset{\sim}{Y} \sim \mathcal{N}(\overline{Y}, N^{-1}\Sigma), \qquad \Sigma|\underset{\sim}{Y} \sim \mathcal{W}^{-1}\{N-1, (NS)\}. \tag{4.44}$$

Using a noninformative improper prior, one would use these results and (4.44) to generate the tth iterate $\xi^{(t)}$ as follows. Suppose individual i, $i = 1, \ldots, N$, has $R_i = r_i$, with corresponding $\bar{r}_i$. Having obtained the tth iterates $Y^{(t)}_{(\bar{r}_i)i}$, say, for each i according to (4.37) (at the tth rather than $(t+1)$th iteration), form for each i

$$Y^{(t)}_i = (Y^T_{(r_i)}, Y^{(t)T}_{(\bar{r}_i)i})^T. \tag{4.45}$$

Write $\underset{\sim}{Y}^{(t)} = \{Y^{(t)}_i, i = 1, \ldots, N\}$.

To obtain $\xi^{(t)}$, first draw random $\Sigma^{(t)}$ from

$$\mathcal{W}^{-1}\{N-1, (NS^{(t)})\}.$$

Many software packages have built-in functions implementing random sampling from an inverse Wishart distribution. Alternatively, it is possible to do this directly, although it could be computationally intensive. Here, one would generate a $\{(N-1) \times K\}$ matrix X, where each of the $N-1$ rows of X are iid draws from

$$\mathcal{N}\{0, (NS^{(t)})^{-1}\},$$

and $S^{(t)}$ is the sample covariance matrix in (4.43) based on $\underset{\sim}{Y}^{(t)}$. Then take

$$\Sigma^{(t)} = (X^T X)^{-1}, \qquad \mu^{(t)} \sim \mathcal{N}(\overline{Y}^{(t)}, N^{-1}\Sigma^{(t)}),$$

where $\overline{Y}^{(t)}$ is the sample mean in (4.43) based on $\underset{\sim}{Y}^{(t)}$.

PROPER IMPUTATION SCHEME: We now summarize ***Step 1*** of the proper multiple imputation procedure using Gibbs sampling based on these results.

Begin with an initial estimator $\xi^{(0)} = (\mu^{(0)}, \Sigma^{(0)})$. One possibility (implemented in SAS `proc mi`, for example) is to use the observed data MLE obtained from the EM algorithm under the assumption of multivariate normality of full data. Set $t = 0$.

Obtain $Y^{(t+1)}_{(\bar{r}_i)i}$ for each $i = 1, \ldots, N$ independently by sampling as in (4.37); that is, generate

$$Y^{(t+1)}_{(\bar{r}_i)i} \sim \mathcal{N}\{\mu^{(t)}_{\bar{r}_i} + \Sigma^{(t)}_{(\bar{r}_i r_i)}(\Sigma^{(t)}_{(r_i r_i)})^{-1}(Y_{(r_i)} - \mu^{(t)}_{r_i}), \Sigma^{(t)}_{(\bar{r}_i \bar{r}_i)} - \Sigma^{(t)}_{(\bar{r}_i r_i)}(\Sigma^{(t)}_{(r_i r_i)})^{-1}\Sigma^{(t)}_{(r_i \bar{r}_i)}\}. \tag{4.46}$$

Form

$$Y^{(t+1)}_i = (Y^T_{(r_i)}, Y^{(t+1)T}_{(\bar{r}_i)i})^T$$

as in (4.45), and construct

$$\overline{Y}^{(t+1)}, S^{(t+1)},$$

the mean and sample covariance matrix in (4.43) based on $\underset{\sim}{Y}^{(t+1)} = \{Y^{(t+1)}_i, i = 1, \ldots, N\}$.

Then obtain $\xi^{(t+1)} = (\mu^{(t+1)}, \Sigma^{(t+1)})$ by first drawing $\Sigma^{(t+1)}$ from

$$\mathcal{W}^{-1}\{N-1, (NS^{(t+1)})\}$$

and then obtaining

$$\mu^{(t+1)} \sim \mathcal{N}(\overline{Y}^{(t+1)}, N^{-1}\Sigma^{(t+1)}).$$

Set $t = t + 1$ and repeat.

One would iterate this process many times to form a (Markov) chain

$$\underset{\sim}{Y}^{(t)} = \{Y_i^{(t)}, i = 1, \ldots, N\} \quad \text{for } t = 1, 2, 3, \ldots$$

The M imputed data sets would then be taken as

$$\underset{\sim}{Y}^{(t)} \quad \text{for } t = s, 2s, \ldots, Ms, \tag{4.47}$$

where s is chosen to be sufficiently large that one feels confident that the Markov chain has stabilized at the stationary distribution.

Multiple imputation ***Steps 2*** and ***3*** would then be carried out based on the M imputed data sets (4.47).

REMARKS:

- This approach to multiple imputation uses multivariate normal model for imputation and thus assumes that all variables in the full data are jointly normally distributed. Even if the variables are all continuous, they may have skewed distributions or otherwise be far from being normally distributed. It may be possible to ***transform*** such variables so that the assumption of approximate normality is more tenable on the transformed scales.
- Most often, complex data sets involve both continuous and ***categorical*** variables, the latter of which may be binary, unordered categorical, or ordered categorical. Clearly, such variables are not normally distributed. One option is just to ignore this complication and hope for the best. As we noted earlier, this may not perform as badly as one might think when the proportion of missing data is not that great. In the next section, we discuss a multiple imputation strategy that attempts to acknowledge and accommodate different types of variables.

- If one in fact believes that the multivariate normal model is ***appropriate*** for the full data, then the full data model $p_Z(z;\theta)$ is a multivariate normal density. In this case, the parameter ξ in the foregoing development and the parameter θ indexing the full data model of interest are one in the same (or are one-to-one transformations of each other). As discussed at the end of Chapter 3 in Section 3.6, inference on θ based on a ***fully Bayesian*** framework is an alternative to the likelihood and multiple imputation methods we have discussed so far. The Gibbs sampling scheme we have presented here in the context of facilitating proper multiple imputation could also be used to implement a fully Bayesian analysis. Specifically, under these conditions, at each iteration t, one obtains a Markov chain

$$\xi^{(t)} = \theta^{(t)}, \quad t = 1, 2, 3, ...$$

  This sequence will converge in distribution to a random θ from the conditional distribution of θ given the observed data; that is, the posterior distribution of θ.

  This suggests that inference on θ could be carried out by running a sufficiently large number s iterations to feel that the Markov chain has stabilized and then taking a large sample $\theta^{(t)}$, $t \geq s$, and using this sample to approximate the posterior distribution. As noted in Section 3.6, one could then use the mean or mode of the sample as an estimator for θ and produce assessments of uncertainty using the sample standard deviation and ***Bayesian credible intervals*** constructed from the sample. A full treatment of the Bayesian approach is beyond our scope here, but this example demonstrates the basic principles of how it might be implemented in general for missing data problems.

4.7 Multivariate imputation by chained equations

As noted above, when the full data comprise both continuous and categorial variables, one strategy is just to proceed as if all were continuous and approximately normal and generate imputations from a multivariate normal distribution. An alternative approach to adopting an imputation model for the entire joint distribution of the full data is to adopt a ***fully conditional specification*** (FCS), also known as ***multivariate imputation by chained equations*** (MICE) (van Buuren, 2007; van Buuren and Groothuis-Oudshoorn, 2011). This multiple imputation approach using ***chained equations*** has gained considerable recent popularity.

This approach is rather ad hoc and involves an algorithm that proceeds as follows.

BASIC IDEA: For simplicity, suppose again the full data are

$$Z = (Y_1, \ldots, Y_K),$$

where each component Y_k of Z is a ***scalar*** that can be observed or missing, and let $R = (R_1, \ldots, R_k)^T$ as usual. The approach can be adapted to the case where each component Z_k of Z is vector-valued by viewing each element of Z_k as a separate variable. Here, each component Y_k can be a real-valued, binary, or unordered or ordered categorical variable.

MICE is an ad hoc, practical approach to generating imputed data sets (so carrying out ***Step 1*** of multiple imputation) based on a set of imputation models, one for each variable with missing values. Assume that the components of Z are ordered in some specific way; different orderings will produce different results.

We first sketch the basic idea of the MICE algorithm, and then provide details on implementation. For a sample of observed data, the algorithm works as follows:

- The algorithm is ***initialized*** by "filling in" all missing values for each individual by random sampling from the observed values. That is, if Y_k is missing for individual i, impute Y_{ik} by sampling at random from the Y_k values for those individuals in the data set for whom Y_k is observed.
- The first variable with missing values for some individuals, Y_1, say, is ***regressed*** on all other variables $Y_2, \ldots, Y_K$, restricting to individuals for whom Y_1 is observed ($R_1 = 1$). Y_1 for individuals for whom Y_1 is missing ($R_1 = 0$) are imputed via simulated draws from the ***corresponding posterior predictive distribution*** of Y_1 (details to be discussed shortly).
- The next variable with missing values for some individuals, Y_2, say, is regressed on all other variables $Y_1, Y_3, \ldots, Y_K$, restricted to individuals for whom Y_2 is observed and using the imputed values for Y_1 obtained in the previous step. Again, Y_2 for individuals for whom Y_2 is missing are imputed by draws from the corresponding posterior predictive distribution of Y_2.
- This scheme continues ***sequentially*** for all other variables with missing values for some individuals. This completes ***one cycle*** of the algorithm.
- This process for all variables is repeated for ***several cycles*** (e.g., 10 to 20) to stabilize the result. This results in a ***single imputed data set***.
- The entire procedure is carried out M times to yield M imputed data sets.

The M imputed data sets can then be used in ***Steps 2*** and ***3*** of multiple imputation.

The term ***chained equations*** thus refers to the successive regression equations (models) used in the sequence for each cycle. The key feature of MICE is the ability to handle ***different variable types*** (continuous, binary, categorical). Namely, the regression models used in the sequence can be specified in accordance with the variable types, as we now demonstrate.

Consider the kth variable with missing values in the sequence, Y_k, so that there are $K-1$ other variables to be used in the regression model for Y_k, some of which are observed on all individuals and others that have been imputed initially or in a previous cycle of the algorithm. For ease of notation, denote these additional variables by $X_1, \ldots, X_{K-1}$. Note that the definition of $X_1, \ldots, X_{K-1}$ changes depending on k.

We describe how, in any cycle, Y_k for each $k \in \{1, \ldots, K\}$ with missing values would be imputed for individuals for whom it is missing when Y_k is continuous, binary, unordered categorical, or ordered categorical. Suppose we are at the tth cycle.

CONTINUOUS Y_K: If Y_k is ***continuous***, the most common choice of model is a ***linear regression model*** assuming Y_k is normally distributed (perhaps on a ***transformed scale***). The approach is similar to that used in the case of ***monotone missingness*** discussed in Section 4.6.

Thus, consider the model

$$Y_k = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{K-1} X_{K-1} + \sigma\epsilon, \quad \epsilon \sim \mathcal{N}(0,1). \tag{4.48}$$

In (4.48) and all other models we specify below, if a variable X_ℓ is categorical, then it would be included in the model via an appropriate dummy variable specification; for brevity, we suppress this in the notation.

Using the data for individuals for whom Y_k is observed; i.e., with i in the set $\{i : R_{ik} = 1\}$, fit this model using OLS (thus assuming observations are independent across i) to obtain $\widehat{\alpha} = (\widehat{\alpha}_0, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_{K-1})^T$, and let

$$\widehat{\sigma}^2 = (N_k - K)^{-1} \sum_{i:R_{ki}=1} (Y_{ik} - \widehat{\alpha}_0 - \widehat{\alpha}_1 X_{i1} - \cdots - \widehat{\alpha}_{K-1} X_{i,K-1})^2,$$

where N_k is the number of individuals with $R_k = 1$.

Then impute Y_k for those individuals for whom it is missing; i.e., individuals i in the set $\{i : R_{ik} = 0\}$, as follows, similar to the procedure used in the monotone missingness case discussed earlier. Obtain

$$\sigma^{2(t)} = \widehat{\sigma}^2(N_k - K)/q,$$

where q is a random draw from a $\chi^2_{N_k-K}$ distribution, and

$$\alpha^{(t)} = \widehat{\alpha} + \sigma^{(t)}(S_k^T S_k)^{-1/2} U,$$

where S_k is the usual design matrix corresponding to the regression model (4.48), and $U \sim \mathcal{N}(0, I_k)$.

Then obtain for each i in the set $\{i : R_{ik} = 0\}$ imputed values

$$Y_{ik}^{(t)} = \alpha_0^{(t)} + \alpha_1^{(t)} X_{i1} + \cdots + \alpha_{K-1}^{(t)} X_{i,K-1} + \sigma^{(t)} u_i,$$

where $u_i \sim \mathcal{N}(0, 1)$.

BINARY Y_k: If Y_k is ***binary***, taking on values 0 or 1, the most common choice of model is the ***logistic regression model***

$$\text{logit}\{\text{pr}(Y_k = 1 | X_1, \ldots, X_{K-1}; \alpha)\} = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{K-1} X_{K-1}, \tag{4.49}$$

where $\text{logit}(p) = \log\{p/(1-p)\}$, and $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_{K-1})^T$.

Let $\widehat{\alpha}$ be the MLE for α obtained using standard logistic regression software to fit (4.49) (e.g., SAS `proc logistic` or R `glm`) based on the data for individuals for whom Y_k is observed; i.e., with i in the set $\{i : R_{ik} = 1\}$. Let $\widehat{\Sigma}$ be the estimated asymptotic sampling covariance matrix for $\widehat{\alpha}$ obtained from the software.

Use a strategy similar to that used in "sort-of proper imputation" in (4.17), and obtain $\alpha^{(t)}$ as a draw from its approximate posterior distribution

$$\mathcal{N}(\widehat{\alpha}, \widehat{\Sigma}).$$

Then impute Y_k for those individuals for whom it is missing; i.e., individuals i in the set $\{i : R_{ik} = 0\}$, by generating $Y_{ik}^{(t)}$ as Bernoulli with success probability π_{ik}, where

$$\pi_{ik} = \text{expit}(\alpha_0^{(t)} + \alpha_1^{(t)} X_{i1} + \cdots + \alpha_{K-1}^{(t)} X_{i,K-1}),$$

where $\text{expit}(u) = e^u/(1 + e^u)$.

UNORDERED CATEGORICAL Y_k: If Y_k is an ***unordered categorical*** variable taking on one of L possible values (that have no natural ordering, such as "red," "blue," "green"), indexed by $\ell = 1, \ldots, L$, then a ***multinomial*** or ***polytomous*** logistic regression model is often used, which specifies

$$\text{pr}(Y_k = \ell | X_1, \ldots, X_{K-1}; \alpha) = \frac{\exp(\alpha_{0\ell} + \alpha_{1\ell} X_1 + \cdots + \alpha_{K-1,\ell} X_{K-1})}{1 + \sum_{\ell'=1}^{L-1} \exp(\alpha_{0\ell'} + \alpha_{1\ell'} X_1 + \cdots + \alpha_{K-1,\ell'} X_{K-1})}, \quad \ell = 1, \ldots, L-1. \tag{4.50}$$

In (4.50), then, the probabilities of being in each of $L-1$ of the categories are modeled, with the Lth category taken as the ***reference*** category, using the fact that the the probabilities for all L categories must sum to one. Thus, in (4.50), there is a $(K \times 1)$ parameter vector $\alpha_\ell = (\alpha_{0\ell}, \ldots, \alpha_{K-1,\ell})^T$ for each $\ell = 1, \ldots, L-1$, for a total of $k(L-1)$ parameters, which we collect as $\alpha = (\alpha_1^T, \ldots, \alpha_{L-1}^T)^T$ $\{K(L-1) \times 1\}$.

Model (4.50) can be fitted by maximum likelihood (based on the multinomial likelihood with these probabilities) using standard software (e.g., SAS `proc logistic` or the `multinom` function in the `nnet` R package). Using the data for individuals for whom Y_k is observed; i.e., with i in the set $\{i : R_{ik} = 1\}$, obtain the MLE $\widehat{\alpha}$ $\{K(L-1) \times 1\}$ and its associated estimated asymptotic covariance matrix $\widehat{\Sigma}$ $\{K(L-1) \times K(L-1)\}$.

As for the binary case, use a strategy similar to that used in "sort-of proper imputation" in (4.17) and obtain $\alpha^{(t)}$ as a draw from its approximate posterior distribution

$$\mathcal{N}(\widehat{\alpha}, \widehat{\Sigma}).$$

Then impute Y_k for those individuals for whom it is missing; i.e., individuals i in the set $\{i : R_{ik} = 0\}$, by generating $Y_{ik}^{(t)}$ as multinomial with probabilities $\pi_{i1}, \ldots, \pi_{iL}$ corresponding to the categories $1, \ldots, L$, where

$$\begin{aligned}
\pi_{i\ell} &= \frac{\exp(\alpha_{0\ell}^{(t)} + \alpha_{1\ell}^{(t)} X_{i1} + \cdots + \alpha_{K-1,\ell}^{(t)} X_{i,K-1})}{1 + \sum_{\ell'=1}^{L-1} \exp(\alpha_{0\ell'}^{(t)} + \alpha_{1\ell'}^{(t)} X_{i1} + \cdots + \alpha_{K-1,\ell'}^{(t)} X_{i,K-1})}, \quad \ell = 1, \ldots, L-1, \\
\pi_{iL} &= \frac{1}{1 + \sum_{\ell'=1}^{L-1} \exp(\alpha_{0\ell'}^{(t)} + \alpha_{1\ell'}^{(t)} X_{i1} + \cdots + \alpha_{K-1,\ell'}^{(t)} X_{i,K-1})}.
\end{aligned}$$

ORDERED CATEGORICAL Y_k: If Y_k is an ***ordered categorical*** or ***ordinal*** variable taking on one of L possible values that have a natural ordering (e.g., "low," "medium," "high"), then a standard model is the ***proportional odds*** or ***ordered logistic model***, the usual form of which is

$$\text{logit}\{\text{pr}(Y_k \leq \ell | X_1, \ldots, X_{K-1}; \alpha)\} = \lambda_\ell - (\alpha_1 X_1 + \cdots + \alpha_{K-1} X_{K-1}), \quad \ell = 1, \ldots, L-1, \tag{4.51}$$

where $\lambda_1 < \cdots < \lambda_{L-1}$, and we define $\alpha = (\lambda_1, \ldots, \lambda_{L-1}, \alpha_1, \ldots, \alpha_{K-1})^T$. Models of the form (4.51) can be fitted by maximum likelihood using standard software (e.g., SAS `proc logistic` or the `polr` function in the R `mass` package).

Using the data for individuals for whom Y_k is observed; i.e., with i in the set $\{i : R_{ik} = 1\}$, obtain the MLE $\widehat{\alpha}$ and its associated estimated asymptotic covariance matrix $\widehat{\Sigma}$. Then, as for the previous models, obtain $\alpha^{(t)}$ as a draw from its approximate posterior distribution

$$\mathcal{N}(\widehat{\alpha}, \widehat{\Sigma}),$$

and impute Y_k for those individuals for whom it is missing; i.e., individuals i in the set $\{i : R_{ik} = 0\}$, by generating $Y_{ik}^{(t)}$ as multinomial with probabilities $\pi_{i1}, \ldots, \pi_{iL}$ corresponding to the ordered categories $1, \ldots, L$, where

$$\pi_{i1} = \frac{\exp\{\lambda_1^{(t)} - (\alpha_1^{(t)} X_{i1} + \cdots + \alpha_{K-1}^{(t)} X_{i,K-1})\}}{1 + \exp\{\lambda_1^{(t)} - (\alpha_1^{(t)} X_{i1} + \cdots + \alpha_{K-1}^{(t)} X_{i,K-1})\}},$$

$$\pi_{i\ell} = \frac{\exp\{\lambda_\ell^{(t)} - (\alpha_1^{(t)} X_{i1} + \cdots + \alpha_{K-1}^{(t)} X_{i,K-1})\}}{1 + \exp\{\lambda_\ell^{(t)} - (\alpha_1^{(t)} X_{i1} + \cdots + \alpha_{K-1}^{(t)} X_{i,K-1})\}} - \frac{\exp\{\lambda_{\ell-1}^{(t)} - (\alpha_1^{(t)} X_{i1} + \cdots + \alpha_{K-1}^{(t)} X_{i,K-1})\}}{1 + \exp\{\lambda_{\ell-1}^{(t)} - (\alpha_1^{(t)} X_{i1} + \cdots + \alpha_{K-1}^{(t)} X_{i,K-1})\}},$$

for $\ell = 2, \ldots, L$,

$$\pi_{iL} = 1 - \frac{\exp\{\lambda_{L-1}^{(t)} - (\alpha_1^{(t)} X_{i1} + \cdots + \alpha_{K-1}^{(t)} X_{i,K-1})\}}{1 + \exp\{\lambda_{L-1}^{(t)} - (\alpha_1^{(t)} X_{i1} + \cdots + \alpha_{K-1}^{(t)} X_{i,K-1})\}}.$$

SUMMARY: The MICE algorithm is implemented in, for example, SAS `proc mi` using the `fcs` option and the R package `mice`. We have sketched the overall approach to imputation here; more detail is given in van Buuren (2007) and van Buuren and Groothuis-Oudshoorn (2011).

REMARK: The MICE/FCS approach is predicated on specification of ***full conditional models***; that is, in our example for each k, models for Y_k as a function of all other variables. Clearly, it is virtually impossible in general, for variables of mixed types, to specify an entire set of such models such that they are all ***compatible***. By this we mean that they all are consistent with a single ***joint distribution*** specification for all variables.

Accordingly, there is no reason to expect, as is the case with MCMC methods based on a compatible conditional models derived from a joint distributional specification, that a sequence of imputed values generated by the algorithm need converge to anything.

We thus emphasize again that the algorithm is ad hoc, although it seems to work well in practice.

4.8 Discussion

Multiple imputation is a popular approach to handling missing data, owing to the fact that "filling in" missing values has great practical appeal. We reiterate that multiple imputation as presented in this chapter is predicated on the validity of the MAR assumption.

The original literature on multiple imputation suggested that M = 3 to 5 imputed data sets are all that may be needed to obtain good results. This was based on a formula derived by Rubin for calculating the ***relative efficiency*** of estimation of θ when using M imputed data sets compared to using an infinite number of imputed data sets under a fraction F of missing information, given by

$$1/(1 + F/M).$$

This formula suggests that. even with 50% of the information in the intended full data missing, one can attain 91% relative efficiency with M = 5 and 95% with M = 10. However, this pertains only to the quality of the multiple imputation estimator for θ ***itself***; it does not address the quality of assessment of uncertainty via Rubin's variance estimator and, for example, confidence intervals and p-values derived using it. Studies of this issue by simulation have shown that a ***much larger number*** M of imputed data sets should be used to obtain accurate estimators of uncertainty. Recommendations of M = 30 to 50 have been made on this basis more recently, and an ad hoc rule of thumb that has been proposed is to take M to be roughly equal to the percentage of missing information.

The somewhat ***circular reasoning*** underlying multiple imputation that we noted earlier; i.e., starting with an efficient estimator to end up with an relatively inefficient estimator, raises the obvious question: "why do multiple imputation?" In addition to the practical appeal, an advantage that has been cited is that, once M imputed data sets have been created, they can be used for multiple purposes and by multiple analysts with different objectives. For example, the data sets could be made publicly available to researchers.

4.9 Additional results

In this section, we examine some of the results in previous sections in more detail. First, we take a closer look at Rubin's variance estimator from a ***Bayesian perspective*** via a heuristic argument. We then present a sketch of the argument leading to the asymptotic normality result for the ***improper imputation estimator*** given in (4.15) as a demonstration of the considerations involved in proving such results.

RUBIN'S VARIANCE ESTIMATOR FROM A BAYESIAN PERSPECTIVE: We begin by observing the ***connection*** between the ***posterior distribution*** of a parameter and the ***asymptotic distribution*** of a consistent and asymptotically normal estimator for the parameter. Recall that, in our discussion of ***sort-of proper imputation***, we noted that making draws from the asymptotic distribution (4.17) of $\widehat{\theta}^{(init)}$ is roughly the same as drawing from the ***posterior distribution*** of θ given the observed data for large N and certain choices of ***prior distribution*** for θ.

Suppose as usual that Z represents the full data, and we assume a full data model $p_Z(z;\theta)$. Suppose we have a sample $\underset{\sim}{Z} = \{Z_1, \ldots, Z_N\}$, assumed to arise from this model.

Suppose further that we put a ***prior*** on θ, $p_\theta(\theta)$, say. Then by ***Bayes' Rule***, the posterior density of θ given the sample of full data $\underset{\sim}{Z}$ is

$$p_{\theta|\underset{\sim}{Z}}(\theta|\underset{\sim}{Z}) = c(\underset{\sim}{Z}) \left\{ \prod_{i=1}^{N} p_Z(Z_i;\theta) \right\} p_\theta(\theta),$$

where $c(\underset{\sim}{Z})$ is a normalizing constant. Taking logarithms yields

$$\log\{p_{\theta|\underset{\sim}{Z}}(\theta|\underset{\sim}{Z})\} = \sum_{i=1}^{N} \log\{p_Z(Z_i;\theta)\} + \log\{p_\theta(\theta)\} + \log\{c(\underset{\sim}{Z}). \tag{4.52}$$

The first term on the right hand side of (4.52) is the full data loglikelihood, maximization of which in θ leads to the ***full data MLE*** $\widehat{\theta}^F$.

As we now demonstrate, under suitable regularity conditions, the ***posterior distribution*** of θ given the full data is ***approximately normal*** with mean equal to the MLE $\widehat{\theta}^F$ and covariance matrix equal to the inverse of the ***full data observed information matrix*** evaluated at $\widehat{\theta}^F$; i.e., $I^F(\underset{\sim}{Z};\widehat{\theta}^F)$.

Expand the full data loglikelihood via a ***Taylor series*** to quadratic terms about the MLE $\widehat{\theta}^F$:

$$\sum_{i=1}^{N} \log\{p_Z(Z_i;\theta)\} \approx \sum_{i=1}^{N} \log\{p_Z(Z_i;\widehat{\theta}^F)\} + \sum_{i=1}^{N} \frac{\partial}{\partial\theta^T} \log\{p_Z(Z_i;\widehat{\theta}^F)\}\,(\theta-\widehat{\theta}^F) \tag{4.53}$$

$$+(1/2)(\theta-\widehat{\theta}^F)^T \left[\sum_{i=1}^{N} \frac{\partial^2}{\partial\theta\partial\theta^T} \log\{p_Z(Z_i;\widehat{\theta}^F)\}\right] (\theta-\widehat{\theta}^F). \tag{4.54}$$

- The first term on the right hand side of (4.53) is a function of the data $\underset{\sim}{Z}$ only, so ***does not*** involve θ.

- The second term on the right hand side of (4.53) involves

$$\sum_{i=1}^{N} S^F(Z_i;\widehat{\theta}^F),$$

  which ***equals zero*** by definition; i.e, $\widehat{\theta}^F$ is the value that solves the ***score equation***.

- In (4.54),

$$-\sum_{i=1}^{N} \frac{\partial^2}{\partial\theta\partial\theta^T} \log\{p_Z(Z_i;\widehat{\theta}^F)\} = I^F(\underset{\sim}{Z};\widehat{\theta}^F).$$

Accordingly, absorbing the first term into a proportionality constant, we have

$$p_{\theta|\underset{\sim}{Z}}(\theta|\underset{\sim}{Z}) \approx d(\underset{\sim}{Z})\,\exp\{-(\theta-\widehat{\theta}^F)^T I^F(\underset{\sim}{Z};\widehat{\theta}^F)(\theta-\widehat{\theta}^F)/2\}\,p_\theta(\theta). \tag{4.55}$$

Note that (4.55) is, up to a proportionality constant, the density as a function of θ of a ***multivariate normal*** with mean $\widehat{\theta}^F$ and covariance matrix $\{I^F(\underset{\sim}{Z};\widehat{\theta}^F)\}^{-1}$.

Note further that, as $N \to \infty$, $I^F(\underset{\sim}{Z};\widehat{\theta}^F)$ gets larger and larger, so that $\{I^F(\underset{\sim}{Z};\widehat{\theta}^F)\}^{-1}$ converges to a ***zero matrix***. This implies that the (posterior) normal density in (4.55) concentrates ***more and more mass*** near $\widehat{\theta}^F$ as $N \to \infty$. Consequently, the prior $p_\theta(\theta)$ plays less and less of a role in determining the density of the posterior distribution of θ as N gets larger.

From this heuristic argument, we see that the ***posterior distribution*** is approximately

$$\mathcal{N}[\,\widehat{\theta}^F, \{I^F(\underset{\sim}{Z};\widehat{\theta}^F)\}^{-1}\,],$$

and thus

$$E(\theta|\underset{\sim}{Z}) \approx \widehat{\theta}^F, \quad \text{var}(\theta|\underset{\sim}{Z}) \approx \{I^F(\underset{\sim}{Z};\widehat{\theta}^F)\}^{-1}. \tag{4.56}$$

By a similar argument, given a sample of ***observed data*** $(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}) = \{(R_i, Z_{(R_i)i}), i = 1, \ldots, N\}$, we can show, analogous to (4.56), the posterior density of θ given the observed sample data, $p_{\theta|\underset{\sim}{R},\underset{\sim}{Z}_{(\underset{\sim}{R})}}(\theta|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})})$, is approximately

$$\mathcal{N}[\,\widehat{\theta}, \{I(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}; \theta)\}^{-1}\,],$$

with

$$E(\theta|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}) \approx \widehat{\theta}, \quad \text{var}(\theta|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}) \approx \{I(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}; \widehat{\theta})\}^{-1}, \tag{4.57}$$

where recall that $\widehat{\theta}$ is the ***observed data MLE***.

Now consider ***proper multiple imputation***, in which we generate M ***artificial full data sets*** $\underset{\sim}{Z}^{(1)}, \ldots, \underset{\sim}{Z}^{(M)}$, where $\underset{\sim}{Z}^{(m)} = \{Z_i^{(m)}, i = 1, \ldots, N\}$, $m = 1, \ldots, M$, from the ***Bayesian posterior predictive distribution***

$$p_{\underset{\sim}{Z}|\underset{\sim}{R},\underset{\sim}{Z}_{(\underset{\sim}{R})}}(\underset{\sim}{z}|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}).$$

For each imputed data set $\underset{\sim}{Z}^{(m)}$, we compute the full data estimator $\widehat{\theta}^{*(m)}$, which, from (4.56), is such that

$$\widehat{\theta}^{*(m)} \approx E(\theta|\underset{\sim}{Z}^{(m)}). \tag{4.58}$$

The multiple imputation estimator is then

$$\widehat{\theta}^{*}_{M} = M^{-1} \sum_{m=1}^{M} \widehat{\theta}^{*(m)} \approx M^{-1} \sum_{m=1}^{M} E(\theta|\underset{\sim}{Z}^{(m)}), \tag{4.59}$$

where we have added the subscript "M" to emphasize explicitly that the estimator (4.59) is based on M imputed data sets.

From (4.57), we have

$$E\{E(\theta|\underset{\sim}{Z}^{(m)})|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}\} = E(\theta|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}) \approx \widehat{\theta}. \tag{4.60}$$

Now consider the multiple imputation estimator $\widehat{\theta}^{*}_{M}$ in (4.59) when $M \to \infty$. Using (4.58) and (4.60), it follows that $\widehat{\theta}^{*}_{M}$ converges in probability to

$$E\{E(\theta|\underset{\sim}{Z}^{(m)})|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}\} = E(\theta|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}),$$

and thus, for large M,

$$\widehat{\theta}^{*}_{M} \approx \widehat{\theta}.$$

Moreover, from a frequentist point of view, we have the large sample result

$$\widehat{\theta} \dot{\sim} \mathcal{N}[\,\theta_0, \{I(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}; \widehat{\theta})\}^{-1}\,],$$

where, from (4.57), we have

$$\{I(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}; \widehat{\theta})\}^{-1} \approx \text{var}(\theta | \underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}) \tag{4.61}$$

By the ***law of total variance***, or ***Eve's Law***, we have that

$$\text{var}(\theta | \underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}) = E\{\text{var}(\theta | \underset{\sim}{Z}) | \underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}\} + \text{var}\{E(\theta | \underset{\sim}{Z}) | \underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}\}. \tag{4.62}$$

It is straightforward that the second term on the right hand side of (4.62), $\text{var}\{E(\theta|\underset{\sim}{Z})|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}\}$, can be estimated ***unbiasedly*** using the M imputations by

$$(M-1)^{-1} \sum_{m=1}^{M} \left\{ E(\theta | \underset{\sim}{Z}^{(m)}) - M^{-1} \sum_{\ell=1}^{M} E(\theta | \underset{\sim}{Z}^{(\ell)}) \right\} \left\{ E(\theta | \underset{\sim}{Z}^{(m)}) - M^{-1} \sum_{\ell=1}^{M} E(\theta | \underset{\sim}{Z}^{(\ell)}) \right\}^{T}. \tag{4.63}$$

Because from (4.58)

$$E(\theta | \underset{\sim}{Z}^{(m)}) \approx \widehat{\theta}^{*(m)},$$

we can approximate (4.63) by

$$(M-1)^{-1} \sum_{m=1}^{M} \left(\widehat{\theta}^{*(m)} - M^{-1} \sum_{\ell=1}^{M} \widehat{\theta}^{*(\ell)} \right) \left(\widehat{\theta}^{*(m)} - M^{-1} \sum_{\ell=1}^{M} \widehat{\theta}^{*(\ell)} \right)^{T}.$$

Similarly, an unbiased estimator for the first term on the right hand side of (4.62), $E\{\text{var}(\theta|\underset{\sim}{Z})|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}\}$, is

$$M^{-1} \sum_{m=1}^{M} \text{var}(\theta | \underset{\sim}{Z}^{(m)}). \tag{4.64}$$

From (4.56),

$$\text{var}(\theta | \underset{\sim}{Z}^{(m)}) \approx \{I^F(\underset{\sim}{Z}; \widehat{\theta}^{*(m)})\}^{-1},$$

suggesting that we can approximate (4.64) by

$$M^{-1} \sum_{m=1}^{M} \{I^F(\underset{\sim}{Z}; \widehat{\theta}^{*(m)})\}^{-1}.$$

Substituting these results into (4.62), and using (4.59), namely, that $\widehat{\theta}^{*}_{M} \approx M^{-1} \sum_{i=1}^{M} \widehat{\theta}^{*(m)}$, an approximate ***unbiased predictor*** for $\text{var}(\theta|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})})$ is then

$$M^{-1} \sum_{m=1}^{M} \{I^F(\underset{\sim}{Z}; \widehat{\theta}^{*(m)})\}^{-1} + (M-1)^{-1} \sum_{m=1}^{M} (\widehat{\theta}^{*(m)} - \widehat{\theta}^{*}_{M})(\widehat{\theta}^{*(m)} - \widehat{\theta}^{*}_{M})^{T}. \tag{4.65}$$

Because of (4.61),

$$\text{var}(\theta|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}) \approx \{I(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}; \widehat{\theta})\}^{-1},$$

(4.65) is an approximate ***unbiased estimator*** for $\{I(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}; \widehat{\theta})\}^{-1}$. Moreover, because $\widehat{\theta}^*_M \approx \widehat{\theta}$ for large M from above, (4.65) is also an estimator for the asymptotic variance of the multiple imputation estimator $\widehat{\theta}^*_M$ as $M \to \infty$.

In fact, (4.65) is very close to ***Rubin's variance estimator*** (4.11). However, we must account for the fact that M, the number of imputations, is ***finite***.

We now take a ***frequentist*** point of view and consider the ***unconditional variance***

$$\text{var}(\widehat{\theta}^*_M) = E\{\text{var}(\widehat{\theta}^*_M|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})})\} + \text{var}\{E(\widehat{\theta}^*_M|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})})\}, \tag{4.66}$$

where the expectations and variances in (4.66) are of course with respect to the ***true distribution*** of the full data, which has density $p_Z(z;\theta_0)$ for some θ_0.

Consider each term in (4.66). Because $\underset{\sim}{Z}^{(m)}$, $m = 1, \ldots, M$, are generated from the "made-up" predictive distribution of $\underset{\sim}{Z}$ given $\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}$, this predictive distribution has nothing to do with θ_0 that generates $\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}$. Thus

$$\begin{aligned} E(\widehat{\theta}^*_M|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}) &= M^{-1}\sum_{m=1}^{M} E(\widehat{\theta}^{*(m)}|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}) \\ &\approx M^{-1}\sum_{i=1}^{M} E\{E(\theta|\underset{\sim}{Z}^{(m)})|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}\} = E(\theta|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}) \approx \widehat{\theta}, \end{aligned}$$

and

$$\text{var}\{E(\widehat{\theta}^*_M|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})})\} \approx \text{var}(\widehat{\theta}),$$

which can be estimated by $\{I(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}; \widehat{\theta})\}^{-1}$.

Likewise,

$$\text{var}(\widehat{\theta}^*_M|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}) = \text{var}\left\{M^{-1}\sum_{i=1}^{M} E(\theta|\underset{\sim}{Z}^{(m)})|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})})\right\} = M^{-1}\text{var}\{E(\theta|Z)|\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})}\}. \tag{4.67}$$

From above, (4.67) comes about from the predictive distribution, having nothing to do with θ_0, and can be estimated by

$$M^{-1}\left\{(M-1)^{-1}\sum_{m=1}^{M}(\widehat{\theta}^{*(m)} - \widehat{\theta}^*_M)(\widehat{\theta}^{*(m)} - \widehat{\theta}^*_M)^T\right\}.$$

From these considerations, (4.66) can be written as

$$E\left\{M^{-1}(M-1)^1\sum_{m=1}^{M}(\widehat{\theta}^{*(m)}-\widehat{\theta}^*_M)(\widehat{\theta}^{*(m)}-\widehat{\theta}^*_M)^T\right\}+\text{var}(\widehat{\theta})$$
$$=E\left\{M^{-1}(M-1)^{-1}\sum_{m=1}^{M}(\widehat{\theta}^{*(m)}-\widehat{\theta}^*_M)(\widehat{\theta}^{*(m)}-\widehat{\theta}^*_M)^T\right\}+\{I(\underset{\sim}{R},\underset{\sim}{Z}_{(\underset{\sim}{R})};\widehat{\theta})\}^{-1}. \quad (4.68)$$

From above, (4.65) is an unbiased estimator for $\{I(\underset{\sim}{R},\underset{\sim}{Z}_{(\underset{\sim}{R})};\widehat{\theta})\}^{-1}$. Taking this together with (4.68), we conclude that an unbiased estimator for $\text{var}(\widehat{\theta}^*_M)$ is given by

$$M^{-1}\sum_{m=1}^{M}\{I^F(\underset{\sim}{Z};\widehat{\theta}^{*(m)})\}^{-1}+\left(\frac{M+1}{M}\right)(M-1)^{-1}\sum_{m=1}^{M}(\widehat{\theta}^{*(m)}-\widehat{\theta}^*_M)(\widehat{\theta}^{*(m)}-\widehat{\theta}^*_M)^T,$$

which is ***Rubin's variance estimator*** (4.11).

HEURISTIC JUSTIFICATION FOR ASYMPTOTIC RESULT FOR IMPROPER IMPUTATION: We now outline the steps in an argument to obtain the asymptotic result for the ***improper imputation estimator*** given in (4.15), repeated here for convenience as

$$N^{1/2}(\widehat{\theta}^{*(improper)}-\theta_0)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\Sigma^{*(improper)}),$$

$$\begin{aligned}\Sigma^{*(improper)} &= \{\mathcal{I}^F(\theta_0)\}^{-1}+\left(\frac{M+1}{M}\right)\{\mathcal{I}^F(\theta_0)\}^{-1}\{\mathcal{I}^F(\theta_0)-\mathcal{I}(\theta_0)\}\{\mathcal{I}^F(\theta_0)\}^{-1}\\ &\quad+\{\mathcal{I}^F(\theta_0)\}^{-1}\{\mathcal{I}^F(\theta_0)-\mathcal{I}(\theta_0)\}\Sigma^{(init)}\{\mathcal{I}^F(\theta_0)-\mathcal{I}(\theta_0)\}\{\mathcal{I}^F(\theta_0)\}^{-1}, \quad (4.69)\end{aligned}$$

and improper imputation is based on an initial estimator $\widehat{\theta}^{(init)}$ with asymptotic covariance matrix $\Sigma^{(init)}(\theta_0)=\Sigma^{(init)}$. For brevity in what follows, we write $\widehat{\theta}^*=\widehat{\theta}^{*(improper)}$.

Here, we assume that the initial estimator $\widehat{\theta}^{(init)}$ for θ is a ***regular, asymptotically linear (RAL) estimator*** ; for example, an M-estimator is ordinarily a RAL estimator. For large N, such an estimator satisfies

$$N^{1/2}(\widehat{\theta}^{(init)}-\theta_0)\approx N^{-1/2}\sum_{i=1}^{N}\varphi^{(init)}(R_i,Z_{(R_i)i}), \quad (4.70)$$

where $\varphi^{(init)}(R,Z_{(R)})$ has ***mean zero*** and is referred to as the ***influence function*** of the estimator $\widehat{\theta}^{(init)}$. It follows that the asymptotic variance of $\widehat{\theta}^{(init)}$ is equal to

$$\text{var}\{\varphi^{(init)}(R,Z_{(R)})\}=E\{\varphi^{(init)}(R,Z_{(R)})\varphi^{(init)}(R,Z_{(R)})^T=\Sigma^{(init)},$$

and

$$N^{1/2}(\widehat{\theta}^{(init)}-\theta_0)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\Sigma^{(init)}).$$

Recall that in improper imputation we draw $Z_i^{(m)}(\widehat{\theta}^{(init)})$ from $p_{Z|R,Z_{(R)}}(z|R_i, Z_{(R_i)i}; \widehat{\theta}^{(init)})$, $i = 1, \ldots, N$, and obtain $\widehat{\theta}^{*(m)}$, $m = 1, \ldots, M$, by solving

$$\sum_{i=1}^{N} S_\theta^F\{Z_i^{(m)}(\widehat{\theta}^{(init)}); \theta\} = 0, \quad m = 1, \ldots, M.$$

We then obtain the improper imputation estimator as $\widehat{\theta}^* = \sum_{i=1}^{M} \widehat{\theta}^{*(m)}$. This estimator is ***asymptotically equivalent*** to the solution to the estimating equation

$$\sum_{i=1}^{N} \left[M^{-1} \sum_{m=1}^{M} S_\theta^F\{Z_i^{(m)}(\widehat{\theta}^{(init)}); \theta\} \right] = 0.$$

By a linear Taylor series expansion, we thus have

$$\begin{aligned} 0 = N^{1/2} \sum_{i=1}^{N} \left[M^{-1} \sum_{m=1}^{M} S_\theta^F\{Z_i^{(m)}(\widehat{\theta}^{(init)}); \widehat{\theta}^*\} \right] &\approx N^{1/2} \sum_{i=1}^{N} \left[M^{-1} \sum_{m=1}^{M} S_\theta^F\{Z_i^{(m)}(\widehat{\theta}^{(init)}); \theta_0\} \right] \\ &+ \left(N^{-1} \sum_{i=1}^{N} \left[M^{-1} \sum_{m=1}^{M} \frac{\partial}{\partial \theta^T} S_\theta^F\{Z_i^{(m)}(\widehat{\theta}^{(init)}); \theta_0\} \right] \right) N^{1/2}(\widehat{\theta}^* - \theta_0). \end{aligned} \tag{4.71}$$

Under regularity conditions, the first term in (4.71) satisfies

$$N^{-1} \sum_{i=1}^{N} \left[M^{-1} \sum_{m=1}^{M} \frac{\partial}{\partial \theta^T} S_\theta^F\{Z_i^{(m)}(\widehat{\theta}^{(init)}); \theta_0\} \right] \xrightarrow{p} E\left[\frac{\partial}{\partial \theta^T} S_\theta^F\{Z_i^{(m)}(\theta_0); \theta_0\} \right] = -\{\mathcal{I}^F(\theta_0)\}^{-1}$$

as $N \to \infty$, because $Z_i^{(m)}(\theta_0)$ has density $p_Z(z; \theta_0)$. We can thus rewrite this as

$$N^{1/2}(\widehat{\theta}^* - \theta_0) \approx \{\mathcal{I}^F(\theta_0)\}^{-1} C_N, \qquad C_N = N^{-1/2} \sum_{i=1}^{N} \left[M^{-1} \sum_{m=1}^{M} S_\theta^F\{Z_i^{(m)}(\widehat{\theta}^{(init)}); \theta_0\} \right]. \tag{4.72}$$

Thus, consider the behavior of C_N in (4.72).

Write C_N as

$$\begin{aligned} C_N = N^{-1/2} \sum_{i=1}^{N} &\left[M^{-1} \sum_{m=1}^{M} S_\theta^F\{Z_i^{(m)}(\theta_0); \theta_0\} \right] \\ &+ \left(N^{-1/2} \sum_{i=1}^{N} \left[M^{-1} \sum_{m=1}^{M} S_\theta^F\{Z_i^{(m)}(\widehat{\theta}^{(init)}); \theta_0\} \right] - N^{-1/2} \sum_{i=1}^{N} \left[M^{-1} \sum_{m=1}^{M} S_\theta^F\{Z_i^{(m)}(\theta_0); \theta_0\} \right] \right). \end{aligned} \tag{4.73}$$

For brevity, define

$$G_i(\theta, \theta_0) = M^{-1} \sum_{m=1}^{M} S_\theta^F\{Z_i^{(m)}(\theta); \theta_0\},$$

so that (4.73) can be written compactly as

$$C_N = N^{-1/2} \sum_{i=1}^{N} G_i(\theta_0, \theta_0) + \left\{ N^{-1/2} \sum_{i=1}^{N} G_i(\widehat{\theta}^{(init)}, \theta_0) - N^{-1/2} \sum_{i=1}^{N} G_i(\theta_0, \theta_0) \right\}. \tag{4.74}$$

Let

$$\lambda(\theta, \theta_0) = E\{G_i(\theta, \theta_0)\},$$

where expectation is with respect to the true distribution of the data. Then the function of θ given by

$$W_N(\theta) = N^{-1/2} \sum_{i=1}^{N} \{G_i(\theta, \theta_0) - \lambda(\theta, \theta_0)\}$$

is a ***centered stochastic process***. Under suitable regularity conditions, the ***theory of empirical processes*** can be used to show that $W_N(\theta)$ converges to a ***mean zero Gaussian process***. This implies ***stochastic equicontinuity*** of the process, under which it can be shown that

$$\begin{aligned} &W_N(\widehat{\theta}^{(init)}) - W_N(\theta_0) \\ &\quad = N^{-1/2} \sum_{i=1}^{N} \{G_i(\widehat{\theta}^{(init)}, \theta_0) - \lambda(\widehat{\theta}^{(init)}, \theta_0)\} - N^{-1/2} \sum_{i=1}^{N} \{G_i(\theta_0, \theta_0) - \lambda(\theta_0, \theta_0)\} \xrightarrow{p} 0. \end{aligned} \tag{4.75}$$

The details of such an argument are beyond our scope here; a sketch can be found in Tsiatis (2006, Section 14.3). Rearranging (4.75), we thus have

$$N^{-1/2} \sum_{i=1}^{N} G_i(\widehat{\theta}^{(init)}, \theta_0) - N^{-1/2} \sum_{i=1}^{N} G_i(\theta_0, \theta_0) \approx N^{1/2} \{\lambda(\widehat{\theta}^{(init)}, \theta_0) - \lambda(\theta_0, \theta_0)\}.$$

It then follows from (4.74) that

$$C_N \approx N^{-1/2} \sum_{i=1}^{N} G_i(\theta_0, \theta_0) + N^{1/2} \{\lambda(\widehat{\theta}^{(init)}, \theta_0) - \lambda(\theta_0, \theta_0)\} \tag{4.76}$$

A Taylor series expansion of the second term on the right hand side of (4.76) yields

$$N^{1/2} \{\lambda(\widehat{\theta}^{(init)}, \theta_0) - \lambda(\theta_0, \theta_0)\} \approx \left\{ \frac{\partial \lambda(\theta, \theta_0)}{\partial \theta^T} \right\}_{\theta=\theta_0} N^{1/2} (\widehat{\theta}^{(init)} - \theta_0).$$

From (4.70), we thus obtain that

$$C_N \approx N^{-1/2} \sum_{i=1}^{N} \left[G_i(\theta_0, \theta_0) + \left\{ \frac{\partial \lambda(\theta, \theta_0)}{\partial \theta^T} \right\}_{\theta=\theta_0} \varphi^{(init)}(R_i, Z_{(R_i)i}) \right]. \tag{4.77}$$

A summand (in brackets) in (4.77) has mean zero. We would thus like to apply the central limit theorem to (4.77) to obtain the limit in distribution of C_N. To do this, we must

(a) Derive $\text{var}\{G_i(\theta_0, \theta_0)\} = E\{G_i(\theta_0, \theta_0) G_i(\theta_0, \theta_0)^T\}$.

(b) Derive $\left\{ \frac{\partial \lambda(\theta, \theta_0)}{\partial \theta^T} \right\}_{\theta=\theta_0}$

(c) Derive $\text{cov}\{G_i(\theta_0, \theta_0), \varphi^{(init)}(R_i, Z_{(R_i)i})\} = E\{G_i(\theta_0, \theta_0) \varphi^{(init)}(R_i, Z_{(R_i)i})^T\}$.

We tackle each of these in turn.

(a) Derive $\text{var}\{G_i(\theta_0, \theta_0)\}$. Recall that

$$G_i(\theta_0, \theta_0) = M^{-1} \sum_{m=1}^{M} S_\theta^F \{Z_i^{(m)}(\theta_0); \theta_0\},$$

and that we start with $\{R_i, Z_{(R_i)i}\}$ and generate $Z_i^{(m)}(\theta_0)$ from the predictive distribution $p_{Z|R,Z_{(R)}}(z|R_i, Z_{(R_i)i})$. Now

$$\begin{aligned}
\text{var}\{G_i(\theta_0, \theta_0)\} &= E[\text{var}\{G_i(\theta_0, \theta_0)\}|R_i, Z_{(R_i)i}] + \text{var}[E\{G_i(\theta_0, \theta_0)|R_i, Z_{(R_i)i}\}] \\
&= E\left(\text{var}\left[M^{-1}\sum_{i=1}^{M} S_\theta^F\{Z_i^{(m)}(\theta_0); \theta_0\} \,\middle|\, R_i, Z_{(R_i)i}\right]\right) + \text{var}\left(E\left[M^{-1}\sum_{i=1}^{M} S_\theta^F\{Z_i^{(m)}(\theta_0); \theta_0\} \,\middle|\, R_i, Z_{(R_i)i}\right]\right) \\
&= E\left[M^{-1}\text{var}\{S_\theta^F(Z_i; \theta_0)|R_i, Z_{(R_i)i}\right] + \text{var}\left[E\{S_\theta^F(Z_i; \theta_0)|R_i, Z_{(R_i)i}\}\right] \qquad (4.78) \\
&= E\left[M^{-1}\text{var}\{S_\theta^F(Z_i; \theta_0)|R_i, Z_{(R_i)i}\}\right] + \text{var}\{S_\theta(R, Z_{(R)})\} \\
&= E\left[M^{-1}\text{var}\{S_\theta^F(Z_i; \theta_0)|R_i, Z_{(R_i)i}\}\right] + \mathcal{I}(\theta_0). \qquad (4.79)
\end{aligned}$$

The equality in (4.78) follows because, conditional on $(R_i, Z_{(R_i)i})$, the $Z_i^{(m)}(\theta_0)$ are independent draws from the predictive distribution $p_{Z|R,Z_{(R)}}(z|R_i, Z_{(R_i)i}; \theta_0)$.

From the argument for the ***missing information principle***, we have

$$\text{var}\{S_\theta^F(Z; \theta_0)\} = E[\text{var}\{S_\theta^F(Z; \theta_0)|R, Z_{(R)}\}] + \text{var}[E\{S_\theta^F(Z; \theta_0)|R, Z_{(R)}\}],$$

which, recognizing that $\text{var}\{S_\theta^F(Z; \theta_0)\} = \mathcal{I}^F(\theta_0)$ and $\text{var}[E\{S_\theta^F(Z; \theta_0)|R, Z_{(R)}\}] = \mathcal{I}(\theta_0)$, yields

$$E[\text{var}\{S_\theta^F(Z; \theta_0|R, Z_{(R)})\}] = \mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0).$$

Substituting this in (4.79), we obtain

$$\text{var}\{G_i(\theta_0, \theta_0)\} = \mathcal{I}(\theta_0) + M^{-1}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}. \qquad (4.80)$$

(b) Derive $\left\{\dfrac{\partial\lambda(\theta, \theta_0)}{\partial\theta^T}\right\}_{\theta=\theta_0}$, where again

$$\lambda(\theta, \theta_0) = E\{G_i(\theta, \theta_0)\} = E\left[S_\theta^F\{Z_i^{(m)}(\theta); \theta_0\}\right]. \qquad (4.81)$$

We can write (4.81) as

$$\lambda(\theta, \theta_0) = E\left(E\left[S_\theta^F\{Z^{(m)}(\theta); \theta_0\} \,\middle|\, R, Z_{(R)}\right]\right), \qquad (4.82)$$

where $Z^{(m)}(\theta)$ is generated from the predictive distribution $p_{Z|R,Z_{(R)}}(z|R, Z_{(R)}; \theta)$.

The inner expectation in (4.82) is thus

$$\int S_\theta^F(z;\theta_0)\, p_{Z|R,Z_{(R)}}(z|r,z_{(r)};\theta)\, d\nu(z_{(\bar{r})}).$$

Because the outer expectation is with respect to the true distribution of $(R, Z_{(R)})$, (4.82) can be written

$$\lambda(\theta,\theta_0) = \int\int S_\theta^F(z;\theta_0)\, p_{Z|R,Z_{(R)}}(z|r,z_{(r)};\theta)\, d\nu(z_{(\bar{r})})\, p_{R,Z_{(R)}}(r,z_{(r)};\theta_0)\, d\nu(z_{(r)})\, d\nu(r) \tag{4.83}$$

We can take the partial derivative of (4.83), interchanging the order of integration and differentiation, to obtain

$$\begin{aligned}\left\{\frac{\partial\lambda(\theta,\theta_0)}{\partial\theta^T}\right\}_{\theta=\theta_0} &= \int\int S_\theta^F(z;\theta_0)\left\{\frac{\partial p_{Z|R,Z_{(R)}}(z|r,z_{(r)};\theta)}{\partial\theta^T}\right\}_{\theta=\theta_0} p_{R,Z_{(R)}}(r,z_{(r)};\theta_0)\, d\nu(z_{(\bar{r})})\, d\nu(z_{(r)})\, d\nu(r)\\ &= \int\int S_\theta^F(z;\theta_0)\,\frac{\partial\log\{p_{Z|R,Z_{(R)}}(z|r,z_{(r)};\theta_0)\}}{\partial\theta^T}\, p_{Z|R,Z_{(R)}}(z|r,z_{(r)};\theta_0)\, p_{R,Z_{(R)}}(r,z_{(r)};\theta_0)\, d\nu(z_{(\bar{r})})\, d\nu(z_{(r)})\, d\nu(r).\end{aligned} \tag{4.84}$$

Because

$$p_{Z|R,Z_{(R)}}(z|r,z_{(r)};\theta) = \frac{p_{R,Z}(r,z;\theta)}{p_{R,Z_{(R)}}(r,z_{(r)};\theta)},$$

we have

$$\log\{p_{Z|R,Z_{(R)}}(z|r,z_{(r)};\theta)\} = \log\{p_{R,Z}(r,z;\theta)\} - \log\{p_{R,Z_{(R)}}(r,z_{(r)};\theta)\},$$

so that

$$\frac{\partial\log\{p_{Z|R,Z_{(R)}}(z|r,z_{(r)};\theta_0)\}}{\partial\theta^T} = S_\theta^F(z;\theta_0)^T - S_\theta(r,z_{(r)};\theta_0)^T. \tag{4.85}$$

Substituting (4.85) into (4.84) yields

$$\left\{\frac{\partial\lambda(\theta,\theta_0)}{\partial\theta^T}\right\}_{\theta=\theta_0} = E\left[S_\theta^F(Z;\theta_0)\{S_\theta^F(Z;\theta_0)^T - S_\theta(R,Z_{(R)};\theta_0)^T\}\right] \tag{4.86}$$

(verify). Now

$$E\{S_\theta^F(Z;\theta_0)S_\theta^F(Z;\theta_0)^T\} = \mathcal{I}^F(\theta_0)$$

and

$$\begin{aligned}E\{S_\theta^F(Z;\theta_0)S_\theta(R,Z_{(R)};\theta_0)^T\} &= E\left[E\{S_\theta^F(Z;\theta_0)S_\theta(R,Z_{(R)};\theta_0)^T|R,Z_{(R)}\}\right]\\ &= E\left[E\{S_\theta^F(Z;\theta_0)|R,Z_{(R)}\}S_\theta(R,Z_{(R)};\theta_0)^T\right] = E\{S_\theta(R,Z_{(R)};\theta_0)S_\theta(R,Z_{(R)};\theta_0)^T\} = \mathcal{I}(\theta_0).\end{aligned}$$

Applying these results to (4.86) leads to

$$\left\{\frac{\partial\lambda(\theta,\theta_0)}{\partial\theta^T}\right\}_{\theta=\theta_0} = \mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0). \tag{4.87}$$

(c) Derive $\text{cov}\{G_i(\theta_0,\theta_0), \varphi^{(init)}(R_i, Z_{(R_i)i})\} = E\{G_i(\theta_0,\theta_0)\varphi^{(init)}(R_i, Z_{(R_i)i})^T\}$. It is straightforward that

$$\begin{aligned} E\{G_i(\theta_0,\theta_0)\varphi^{(init)}(R_i, Z_{(R_i)i})^T\} &= E\left[E\{G_i(\theta_0,\theta_0)\varphi^{(init)}(R_i, Z_{(R_i)i})^T | R_i, Z_{(R_i)i}\}\right] \\ &= E\{S_\theta(R_i, Z_{(R_i)i};\theta_0)\varphi^{(init)}(R_i, Z_{(R_i)i})^T\} = I_p, \end{aligned} \tag{4.88}$$

where I_p is a $(p \times p)$ identity matrix, and p is the dimension of θ. The final equality follows by a well-known property of ***influence functions*** of RAL estimators, which is proved in Theorem 3.2 of Tsiatis (2006).

Having demonstrated (a)-(c), we can now deduce the final result. Using (4.80), (4.87), and (4.88), we have that

$$C_N \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_C), \tag{4.89}$$

$$\Sigma_C = \mathcal{I}(\theta_0) + M^{-1}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\} + \{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}\Sigma^{(init)}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}^T + 2\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}.$$

Thus, from (4.72),

$$N^{1/2}(\widehat{\theta}^* - \theta_0) \approx \{\mathcal{I}^F(\theta_0)\}^{-1} C_N \xrightarrow{\mathcal{L}} \mathcal{N}\left[0, \{\mathcal{I}^F(\theta_0)\}^{-1}\Sigma_C\{\mathcal{I}^F(\theta_0)\}^{-1}\right],$$

from which we obtain the final result that

$$N^{1/2}(\widehat{\theta}^{*(improper)} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma^{*(improper)}), \tag{4.90}$$

because, by straightforward algebra (try it),

$$\begin{aligned} &\{\mathcal{I}^F(\theta_0)\}^{-1}\Sigma_C\{\mathcal{I}^F(\theta_0)\}^{-1} \\ &\quad= \{\mathcal{I}^F(\theta_0)\}^{-1}\Big[\mathcal{I}(\theta_0) + M^{-1}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\} \\ &\qquad + \{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}\Sigma^{(init)}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}^T + 2\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}\Big]\{\mathcal{I}^F(\theta_0)\}^{-1} \\ &\quad= \{\mathcal{I}^F(\theta_0)\}^{-1} + \left(\frac{M+1}{M}\right)\{\mathcal{I}^F(\theta_0)\}^{-1}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}\{\mathcal{I}^F(\theta_0)\}^{-1} \\ &\qquad + \{\mathcal{I}^F(\theta_0)\}^{-1}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}\Sigma^{(init)}\{\mathcal{I}^F(\theta_0) - \mathcal{I}(\theta_0)\}\{\mathcal{I}^F(\theta_0)\}^{-1} \\ &\quad= \Sigma^{*(improper)}. \end{aligned}$$